2009

Analysis of Machine Learning Based Methods for Identifying MicroRNA Precursors

Steve Ikeoka
San Jose State University

ANALYSIS OF MACHINE LEARNING BASED METHODS FOR IDENTIFYING MICRORNA PRECURSORS

A Writing Project

Presented to

The Faculty of the Department of Computer Science

San José State University

In Partial Fulfillment

of the Requirements for the Degree

Master of Science

by

Steve Ikeoka

December 2009

APPROVED FOR THE DEPARTMENT OF COMPUTER SCIENCE

Dr. Sami Khuri, Department of Computer Science Date

Dr. Mark Stamp, Department of Computer Science Date

Dr. Robert Chun, Department of Computer Science Date

APPROVED FOR THE UNIVERSITY

Associate Dean Office of Graduate Studies and Research Date

ABSTRACT

ANALYSIS OF MACHINE LEARNING BASED METHODS
FOR IDENTIFYING MICRORNA PRECURSORS

By Steve Ikeoka

MicroRNAs are a type of non-coding RNA that were discovered less than a decade ago but are now known to be incredibly important in regulating gene expression despite their small size. However, due to their small size, and several other limiting factors, experimental procedures have had limited success in discovering new microRNAs. Computational methods are therefore vital to discovering novel microRNAs. Many different approaches have been used to scan genomic sequences for novel microRNAs with varying degrees of success. This work provides an overview of these computational methods, focusing particularly on those methods based on machine learning techniques. The results of experiments performed on several of the machine learning based microRNA detectors are provided along with an analysis of their performance.

ACKNOWLEDGEMENTS

I would like to thank my advisor Dr. Sami Khuri for introducing me to the field of bioinformatics and for his guidance and support throughout the project. I would also like to thank Dr. Mark Stamp and Dr. Robert Chun for participating in my committee. Finally, I want to thank my parents for teaching me the value of a good education and hard work and supporting me in achieving my educational goals.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. Introduction

1.1. Introduction to Bioinformatics

Bioinformatics is the application of mathematics and computer science to solve problems in the field of molecular biology. The field of molecular biology has been advancing rapidly since James Watson and Francis Crick discovered the molecular structure of DNA in 1953. With this advancement in molecular biology, the amount of experimental data generated by laboratories around the world has also increased tremendously. The size and complexity of this information has created new problems since biologists now need help from computers to use all of this data effectively [19].

There are many different examples of computational techniques being used to help with solving biological problems. One such example is gene prediction. Gene prediction involves identifying where the genes are in a given genomic DNA sequence. In the case of protein-coding genes, the gene is transcribed from DNA into a messenger RNA (mRNA) molecule and the mRNA is translated into a protein. For eukaryotic organisms, the mRNA will undergo additional processing, such as splicing of introns, before being translated. There are many computer programs available for gene prediction, such as ORPHEUS and GLIMMER for prokaryotes and GenScan for eukaryotes. Gene prediction programs such as these have been vital in discovering new genes and understanding their functions [31].

1.2. Introduction to Non-Coding RNAs (ncRNAs)

In addition to protein-coding genes, there are also many genes for which the functional product is RNA. Functional RNAs which are not translated into a protein are known as non-coding RNAs (ncRNAs). There are many examples of ncRNAs, most notably transfer RNA (tRNA) and ribosomal RNA (rRNA). Many ncRNA families have secondary structures which are highly conserved across many species and computational methods for detecting ncRNA genes rely on this property. It is advantageous to identify those ncRNA genes first when annotating a newly sequence genome because they are generally easier to identify than protein-coding genes [31].

1.3. Project Overview

My project focuses on a particular type of ncRNA called microRNA (miRNA). Chapter 2 provides a background on the formation, function and importance of miRNAs and explains two research problems involving miRNAs. Chapter 3 describes some of the online resources for storing miRNA data. Chapter 4 describes some of the computational approaches for detecting miRNAs, with an in-depth explanation of several machine learning based miRNA methods. Chapter 5 describes the software and data that I used in my experiments. Chapter 6 discusses the statistics that I used to analyze my results. Chapter 7 provides the results of my experiments and a discussion about these results is provided in Chapter 8. The conclusions from my work are provided in Chapter 9.

2. MicroRNA Background

MicroRNAs represent a large family of small ncRNAs. The first of what are now known as miRNAs, lin-4, was actually discovered in 1993 but it was originally thought to be some sort of a genetic quirk. It wasn't until 2001 that researchers discovered that this type of small RNA was widespread in animals and the term 'microRNA' was introduced. The first experiments involving the cloning and identification of miRNAs in plants were reported in mid-2002, "demonstrating that miRNAs are a fundamental feature of multicellular eukaryotic life" [13].

2.1. Formation of miRNA

Figure 1 illustrates the formation of miRNAs. There are two major pathways through which miRNAs are formed. In the first case, the miRNA is encoded by a gene which is transcribed to form the primary miRNA (pri-miRNA). As shown in the top left of Figure 1, it is also possible for miRNA genes to exist in a cluster that is transcribed into a single pri-miRNA containing multiple RNAs. An enzyme called Drosha processes the pri-miRNA by cleaving out the stem-loop structures which become the precursor miRNAs (pre-miRNA). In the second pathway, the pre-miRNA is actually contained inside of a special type of intron of a protein-coding gene, called a mirtron. When the gene is expressed, the mirtron is spliced out of the messenger RNA molecule and a special enzyme extracts the pre-miRNA from the mirtron [15].

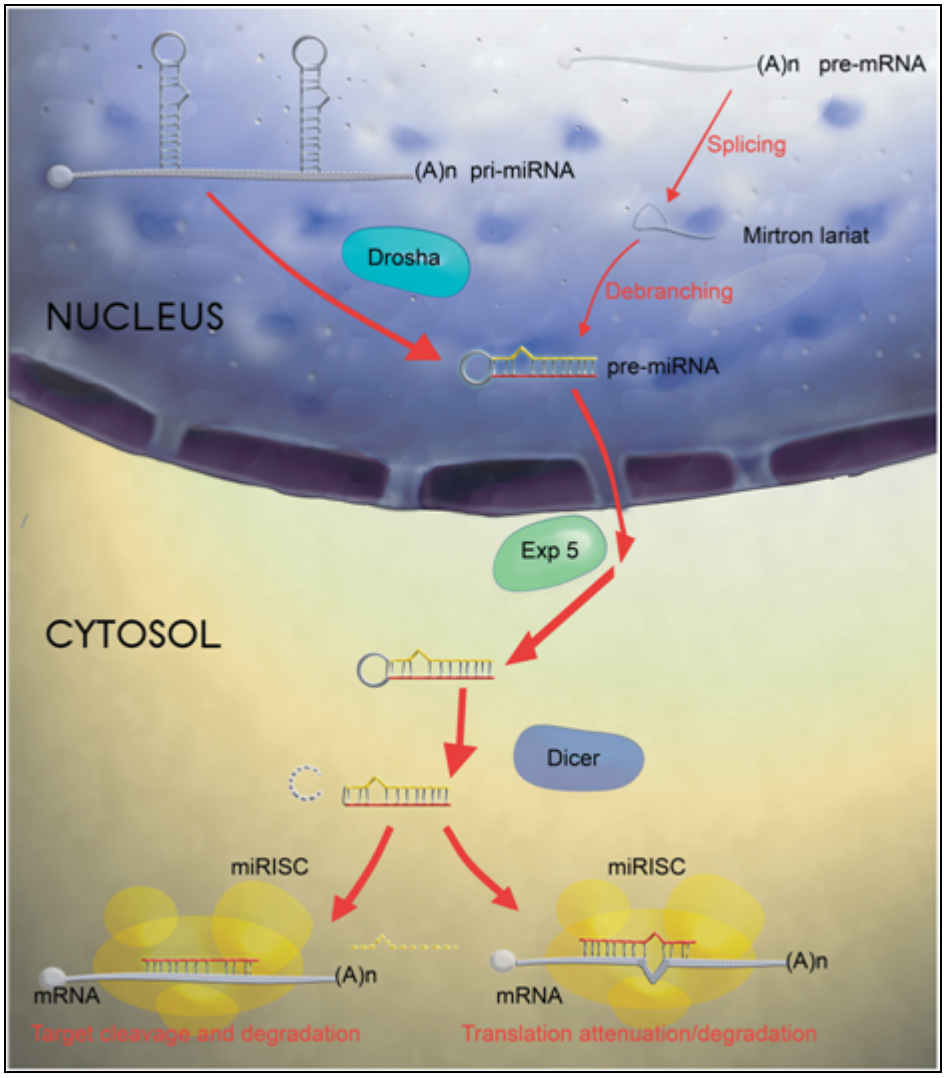

Figure 1. The formation of miRNA [15]

The pre-miRNA, which is about 70 nucleotides long and has a stem-loop secondary structure (Figure 2), is then transported from the nucleus to the cytoplasm by a complex of Exportin 5 (Exp 5) and Ran-GTP proteins. In the cytoplasm, the enzyme Dicer processes the pre-miRNA to generate the mature miRNA, which is about 22 nucleotides long. The mature miRNA is finally integrated into the miRNA-induced silencing complex (miRISC) [15].

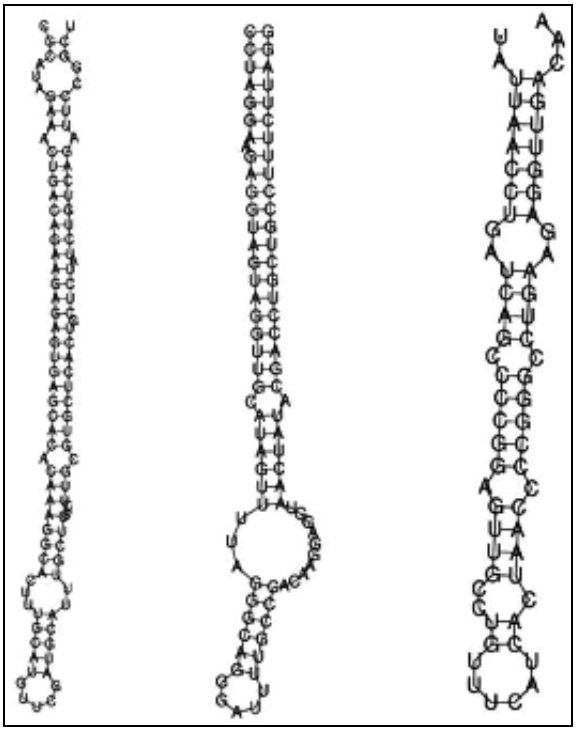

Figure 2. The pre-miRNA stem-loop secondary structure [5]

2.2. Function of miRNA

The function of miRNA is regulating gene expression. As shown at the bottom of Figure 1, the mature miRNA becomes a part of the miRISC complex which binds to a target mRNA molecule to either degrade the mRNA or repress its translation depending on how the mature miRNA complements the mRNA target site. Figure 3 shows that when the mature miRNA perfectly complements the mRNA target, the mRNA is degraded. More commonly, the mature miRNA and the mRNA are not a perfect complement, which results in protein translation being repressed.

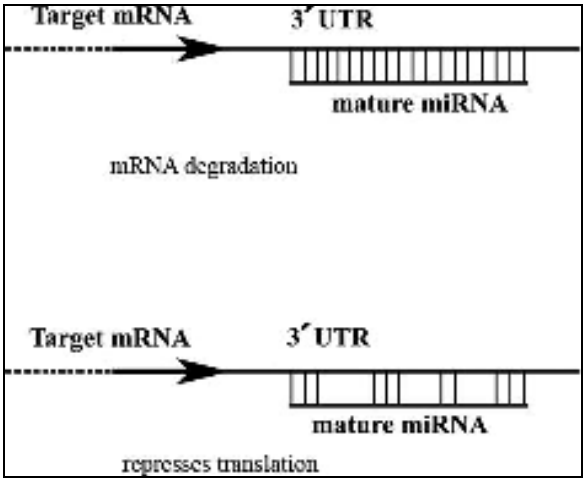

Figure 3. The function of mature miRNA [5]

2.2.1. Links to Diseases

MicroRNAs have been linked to certain types of cancers. One particular miRNA, oncomir-1, has been found to be expressed at abnormally high levels in B cell lymphomas. In one experiment, mice that were engineered to overexpress this miRNA developed tumors as early as two months old whereas mice with the normal miRNA gene developed tumors between six to nine months old [27]. Abnormally high amount of other miRNAs have been linked to malignant tumors in the liver, breast, colon and lymph nodes. Because of this link, researchers are currently studying the possibility of using miRNAs and other small ncRNAs to diagnose the origins of tumors by measuring the patterns of expression levels of different RNAs. Knowing where the cancer originated is vital to appropriate treatment when the tumors have spread throughout the body but reliable cancer biomarkers are currently lacking.

2.2.2. Pharmaceutical Benefits

Research into the effects of these small RNAs suggests that "microRNAs appear to function much like a set of biological master keys" [27]. A single miRNA is able to regulate the mRNA sequences of many different genes because, as shown in Figure 3, the miRNA can bind to the mRNA even without perfect base pairing. Despite their dramatic effects, the structures of miRNAs are relatively simple, which is why many researchers feel that miRNAs have "great potential for their use as pharmaceuticals" [27].

Many current drugs are designed to interact with proteins but it is difficult to design such drugs because the three-dimensional structure of proteins is very complex. On the other hand, RNAs have a very simple structure so it is relatively straightforward to design a drug to interact with a particular RNA. It is estimated that the human body contains about 700 different miRNAs so designing drugs to regulate gene expression by manipulating miRNAs has significant potential [27].

2.3. Research Problems

There are two major research problems involving miRNAs. The first problem is detecting the miRNA genes, which is the focus of this work, and the second problem is predicting the location of miRNA targets. This section will provide a description of both of these problems.

2.3.1. Gene Detection

Finding miRNA through experimental approaches is very difficult due to a number of factors that limit the effectiveness of conventional genetic techniques, such as direct cloning and the use of mutagenesis. These limiting factors include "the short length of miRNAs and their ability to act redundantly or to have only a subtle phenotypical impact" and "miRNAs that have very low expression levels or that are expressed only in specific conditions and cell types" [15]. Deep-sequencing techniques, which require extensive computational analysis, have had some success in overcoming these limitations but it is clear that sophisticated computational approaches are vital to finding novel miRNAs.

2.3.2. Target Prediction

MicroRNAs are involved in regulating gene expression and as shown in Figure 3, the two major functions of miRNAs are degrading the mRNA or repressing its translation and which method is used depends on the complementarity between the miRNA and the target location in the mRNA. It is also common in animals that a particular miRNA will have multiple targets on the same mRNA or that multiple miRNAs could target the same mRNA. Understanding exactly how miRNAs regulate gene expression is vital to the field of miRNA research. Additional information about computational approaches to target prediction can be found in [5] and [15].

3. MicroRNA Databases

With the number of known and predicted miRNAs and their targets increasing rapidly, computer databases developed to handle this type of data have been essential to storing and organizing all of these data so that it can be utilized by researchers around the world. Table 1 lists some of the available online database resources for miRNAs and their targets along with their URL and a very brief description of the database. There are many other resources available, some which specialize in miRNAs and other that are general RNA resources like Rfam. This section will provide more background on two of the more popular resources, Rfam and miRBase.

Table 1. List of miRNA databases [5]

Name	URL	Description
Rfam	http://rfam.sanger.ac.uk/ and http://rfam.janelia.org/	Annotation and alignments of RNA families
miRBase	http://microrna.sanger.ac.uk/	Published miRNA sequences, predicted miRNA targets
miRNAMap	http://mirnamap.mbc.nctu.edu.tw/	Known miRNAs, experimental miRNA targets, expression profiles
microRNA.org	http://www.microrna.org/	miRNA targets and expression profiles
TarBase	http://diana.cslab.ece.ntua.gr/	Database of experimentally supported miRNA targets
MirGen	http://diana.cslab.ece.ntua.gr/	Integrated database of animal miRNAs and predicted targets
Argonaute	http://www.ma.uni-heidelberg.de/apps/zmf/argonaute/	Mammalian miRNAs and their known or predicted targets

3.1. Rfam

Rfam [7] is a database of RNA families that represents each family by a multiple sequence alignment, consensus secondary structure, and covariance model (CM). The latest version, which was released in December 2008, contains 1371 RNA families. The CM is a slightly more complicated version of a hidden Markov model (HMM) that is designed to simultaneously model both the sequence and structure of the RNA. Using the INFERNAL package, the CM can be used to search genomes or DNA sequence databases for homologs of a known RNA family. An example of the consensus secondary structure obtained from Rfam is the secondary structure for mir-1302 family shown in Figure 4.

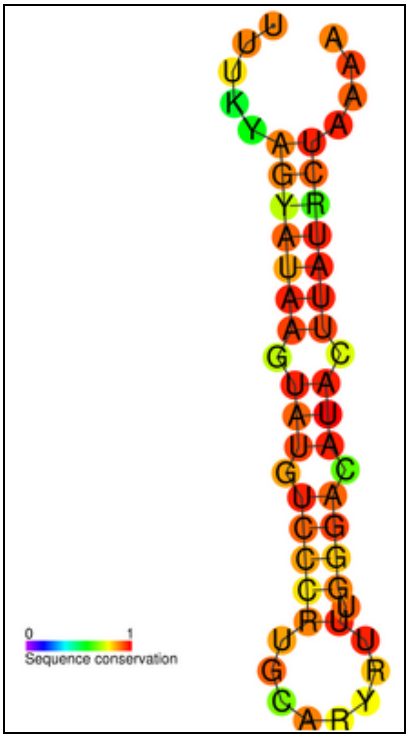

Figure 4. mir-1302 secondary structure from Rfam

Figure 5. miRBase entry for hsa-mir-1302-1

3.2. miRBase

The miRBase [8] repository consists of three separate tools. The miRBase Registry is a confidential service that provides researchers with unique names for their novel miRNA genes before publishing their results. The miRBase Targets database is a resource that provides the predicted targets of all published animal miRNAs. The current

release (version 5) contains predicted targets for all miRNAs in 24 species. The miRBase Sequence Database is a database of all published miRNAs and their annotations. Release 13.0 of the database, which was released in March 2009, contains 9539 miRNA precursors in 103 species. Figure 5 shows the entry from the miRBase Sequence Database for the miRNA hsa-mir-1302-1. The entry contains information about the stem-loop sequence (the pre-miRNA), the mature miRNA sequence and references to the articles where the sequence was published.

4. MicroRNA Detectors

Conventional gene predictors rely on "the characteristic statistical properties of coding regions" to find genes but these techniques do not work for finding miRNAs since non-coding genes do not get translated into a protein and therefore do not exhibit these same properties [15]. Additionally, it is very difficult to obtain an evolutionary model for miRNAs because the precursor and mature miRNA sequences are so short. The lack of a clear evolutionary model limits the use of homology-based searches.

Computational approaches to finding miRNA genes rely on three known properties of miRNAs. The first property is that miRNAs from the same gene family have a very high sequence similarity. The second property is that pre-miRNAs, which are about 70 nucleotides long, form a stable stem-loop secondary structure. The third

property is that the mature mRNA, which is about 22 nucleotides long, is located in the stem region of the pre-miRNA instead of the loop region.

Detecting miRNAs is more challenging in plant genomes than in animals because plant pre-miRNAs have less sequence conservation and a more variable hairpin structure length compared to animal pre-miRNAs. This has justified different computational approaches to finding miRNAs in animals and plants and this work will focus only on detecting animal miRNAs.

The current computational methods to identifying miRNA genes in animals can be categorized in five general approaches: filter-based, homology-based, target-centered, machine learning and mixed approaches [15].

4.1. Filter-Based Approaches

The earliest methods for finding miRNA gene were based on identifying a small number of conserved stem-loop candidates. These filter-based approaches consist of four basic steps: identifying the initial candidate set, restricting the candidates based on structure criteria, further restriction using conservation criteria and, in some cases, using additional filters.

A simple filter-based procedure named MiRscan is described in [5]. It uses a 110 nucleotide long sliding window and folds the window with the RNA folding algorithm RNAFold to identify stem-loop structures with a minimum length and minimum free energy. These conserved stem-loops are considered to be the potential pre-miRNAs. A 21 nucleotide long sliding window is then used to scan each of the potential pre-miRNAs for sequences that have sequence similarity to known miRNAs.

Phylogenetic shadowing is an approach to cross-species sequence comparison that allows for "unambiguous sequence alignments and accurate conservation determination at single nucleotide resolution level" [2]. This approach has been applied to consider conservation in the sequence surrounding the miRNA precursor region while searching for mammalian miRNAs. Phylogenetic shadowing revealed a distinctive drop in conservation in the sequences immediately adjacent to the miRNA stem-loops, which was used to create a characteristic conservation profile to predict novel miRNAs.

These two methods, along with many other filter-based methods, have been able to recover a vast majority of the known miRNAs. Unfortunately, these methods "have failed to produce a set of rules capable of recovering all known miRNAs without leading to too many false positives" [15]. Another problem with these methods is that they are not able to identify non-conserved miRNA candidates because their accuracy relies heavily on the conservation criteria.

4.2. Homology-Based Searches

In biology, homologous genes refer to genes that have similar properties due to some shared evolutionary ancestor. Many homology-based searches rely only on sequence conservation while more sophisticated methods have incorporated structure conservation to increase the sensitivity of the search since RNA structure is generally more conserved than its sequence. Two uses of homology-based methods are "to scan newly sequenced genomes for homologues of known miRNA, or to futher saturate miRNA gene predictions in previously studied genomes" [15].

A profile-based method that exploits both structure and sequence conservation is described in [14]. This method relies on a program called ERPIN [6] that uses multiple sequence alignments to construct profiles that represents both the primary and secondary structures of the RNA family. The authors reported that their profile-based detector discovered 17% more novel miRNA candidates compared to a BLAST search. This suggests that methods that rely only on sequence similarity and methods that combine sequence and structure similarity should be combined to increase the number of predicted miRNA candidates.

Another homology-based approach called miRAlign, which relies on sequence and structure alignments, is described in [24]. The advantage of miRAlign's structure alignment approach compared to the previously described profile-based method is that

constructing the ERPIN profile requires a large number of known family members but miRAlign uses a position independent scoring matrix which can query a single miRNA in the homology search.

4.3. Target-Centered Approaches

An innovative approach based on comparative genomics was applied to finding miRNA genes in [28]. The authors constructed alignments of the 3'-UTRs of human, mouse, rat and dog genomes and used the alignments to discover highly conserved motifs that could be potential miRNA targets. They then searched for conserved regions in the four mammalian genomes complementary to these short motifs. An RNA folding program was used on the conserved site and the flanking sequences to identify potential stem-loop structures. This target-centered approach was able to recover several known miRNAs and well as discover new miRNAs. This approach relies on finding highly conserved motifs in the 3'-UTRs so it will not be able to discover all possible miRNA targets but the advantage of this approach is that it does not rely too heavily on assumptions about pre-miRNA secondary structures.

4.4. Mixed Approaches

Mixed approaches attempt to combine high-throughput experimental procedures with computational methods. There are two basic approaches that are used. The first

approach uses computational methods to generate a large number of potential candidate pre-miRNAs and uses experimental methods to verify the actual miRNAs from the false positives. The other approach uses experimental cloning techniques to generate a large number of small RNA candidates and used computational methods to determine their potential of forming a stem-loop structure.

4.5. Machine Learning Methods

Machine learning methods attempt to generalize between a positive training set consisting of known miRNAs and a negative training set which consist of stem-loop structures that are assumed to not be pre-miRNAs. Section 4.5.1 presents a machine learning method based on naïve Bayes classifiers. Section 4.5.2 presents two methods based on hidden Markov models (HMM). Support vector machine (SVM) based methods, which are the most common machine learning method for miRNA prediction, are presented next in Section 4.5.3. Finally, several methods which rely on other machine learning techniques are briefly presented in Section 4.5.4.

4.5.1 Naïve Bayes Classification

BayesMiRNAfind is a miRNA gene prediction program that utilizes the naïve Bayes classifier [30]. Compared to other machine learning methods, naïve Bayes is a relatively simple and easy to implement classification model that assumes conditional

independence of the features given the class but naïve Bayes models still tend to perform well often. The model used by BayesMiRNAfind is generated from a weighted combination of the feature vector consisting of 62 secondary structure features, such as the number of bulges or the number of loops, and 12 sequence-based features. The classification model was trained on a set of all known miRNAs from multiple species. The entire pipeline for the program is shown in Figure 6. A 110 nucleotide sliding window is run through an RNA folding algorithm to extract potential stem loop structures. A 21 nucleotide long sliding window is used to find potential mature miRNAs within each candidate stem-loop structure and the naïve Bayes classifier is used to find the highest scoring mature miRNA candidate within each stem-loop. An appropriate threshold is then applied to reduce the number of false positives. A conservation filter is also applied which retains only the sequences "which are highly conserved with respect to the reference genome" [30].

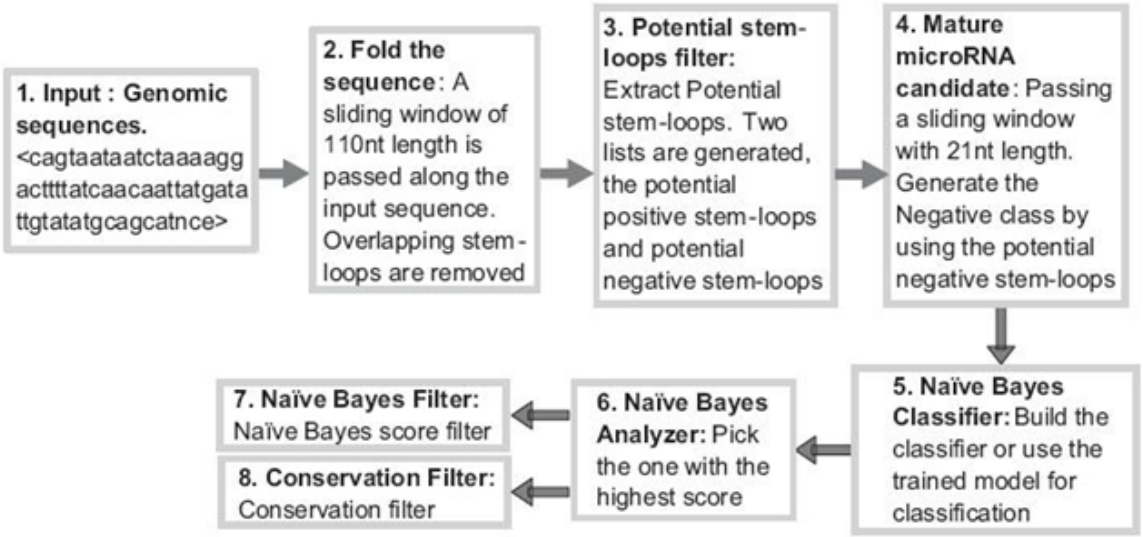

Figure 6. Pipeline of Naïve Bayes algorithm [30]

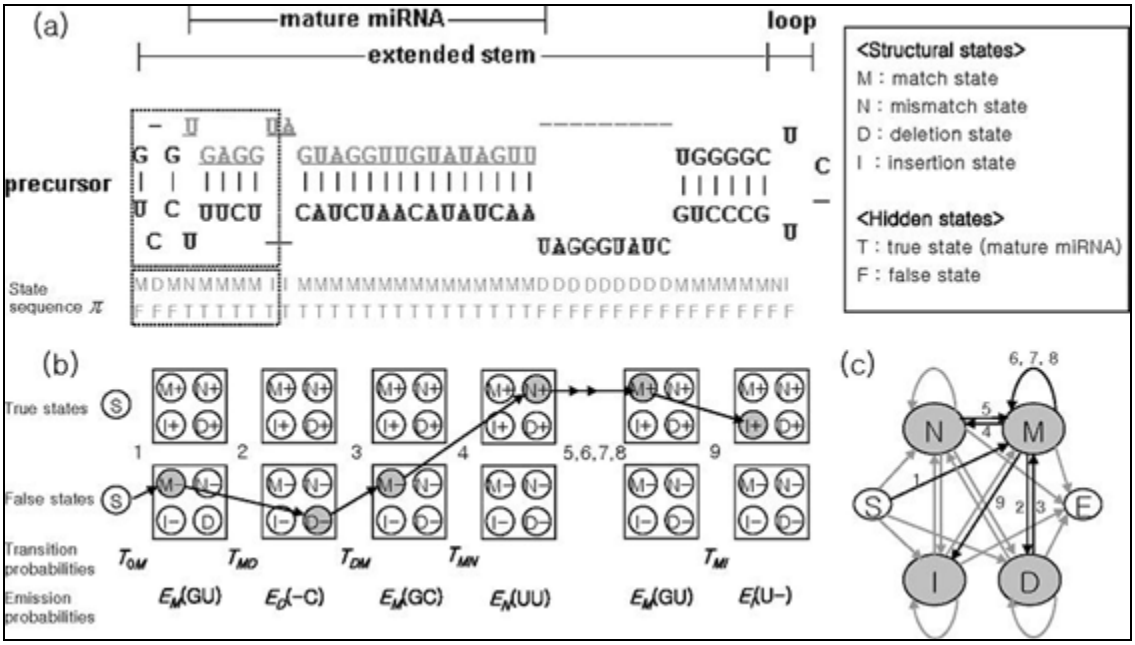

Figure 7. ProMiR representation of pre-miRNA sequence and structure [16]

4.5.2 Hidden Markov Model

ProMiR uses a probabilistic co-learning model that combines characteristics of both the sequence and structure of pre-miRNAs [16]. It uses a paired hidden Markov model (HMM) to search for the Drosha cleavage site of miRNA genes and is able to predict conserved as well as non-conserved miRNAs. MicroRNA precursors form a stem-loop structure which can be represented by a sequence of base pairs where the state of each base pair depends on its base pairing status, which could be a match, mismatch, deletion or insertion. Using this representation, each position consists of a structural state, which is what was just described, and a hidden state, which classifies the position as being either inside or outside the mature miRNA region (Figure 7a, b). All of the possible state transitions in the HMM are shown in Figure 7c. ProMiR II is an enhanced

version that incorporates additional filtering criteria to allow "for low- or high-stringency prediction of conserved and non-conserved miRNA genes" [17].

Another HMM-based miRNA predictor, called the microRNA region inference mechanism (miRRim), was designed to detect highly conserved miRNAs in mammals [23]. In this method, the miRNA and the immediately flanking sequences are represented by a sequence of vectors consisting of five evolutionary and secondary structure features. The first feature is the conservation score (CS) which is a measure of conservation based on a multiple alignment. The second feature is the Z-score, which is calculated based on the minimum free energy (MFE) of the candidate region. The remaining three features are the left and right stem potentials (P^L and P^R), which represent the probability of the position being the left and rights sides of the base pair, and the loop potential (V'), which represents the probability of the position being in the loop of the stem-loop structure. Figure 8 shows the feature vector where the values at each position in the sequence from all of the training samples were averaged.

In order to distinguish between miRNA regions and non-miRNA regions, four HMMs were constructed. One HMM represented the miRNA regions and the other three HMMs represented the non-miRNA regions based on the level of conservation, either nonconserved, moderately conserved, or highly conserved. The final HMM is simply the four HMMs connected together where the transition probability between the miRNA and non-miRNA region HMMs, τ, controls the stringency of the predictions (Figure 9).

Decreasing the τ value increases the stringency, resulting in fewer false positives, while increasing the τ value decreases the stringency, resulting in more false positives.

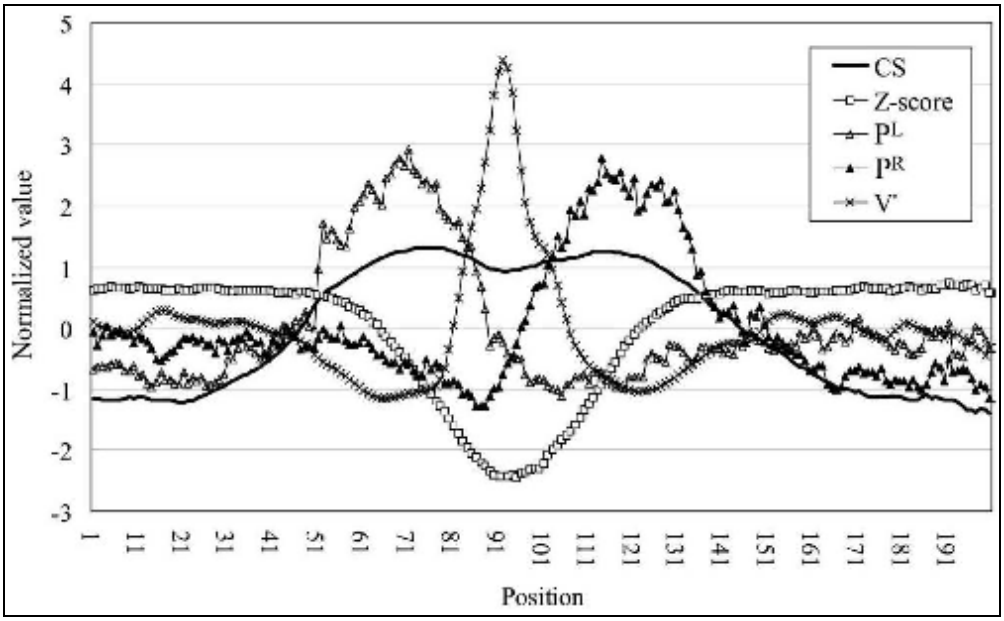

Figure 8. Average values of miRRim feature vector [23]

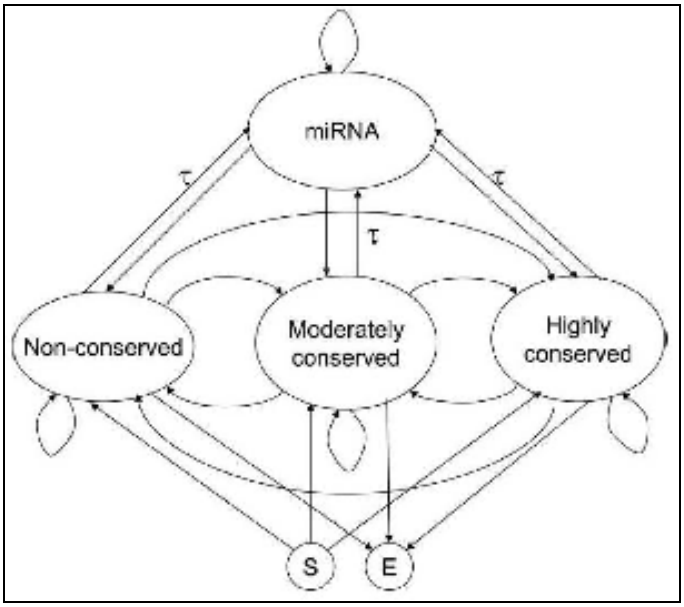

Figure 9. miRRim HMM structure [23]

4.5.3 Support Vector Machine

Table 2. List of some SVM-based miRNA predictors

SVM Classifier	Total # of Features	Sequence Composition	Topological Properties	Thermodynamic Stability	Entropy Measures
Triplet-SVM [29]	32	X			
[1] miRNA SVM [9]	18	X	X		
[2] mirCoS [21]	12	X	X	X	
RNAmicro [10]	12	X	X	X	
miR-abela [20]	40	X	X	X	
miPred [18]	29	X	X	X	X
microPred [1]	48	X	X	X	X
MiRFinder [11]	18		X	X	

[1] [9] uses a preprocessor SVM and a classification SVM
[2] [21] consists of 3 different SVMs applied sequentially

The primary objective of the support vector machines (SVM) is to separate "a set of complex feature vectors into binary labeled classes" and one of the advantages of the SVM is that they are capable of dealing "easily with multi-dimensional data sets that can be noisy or redundant (non-informative or highly correlated)" [18]. SVMs are the most popular machine learning method used to predict miRNA genes. Table 2, which is partially based on information compiled by Mendes et al. [15], shows a summary of eight different SVM classifiers, although there are many others that have been developed for this problem. As shown in Table 2, the total number of features that different SVMs use to classify sequences varies significantly. Three major sets of features that are used by SVMs for predicting miRNAs are sequence composition, topological properties and thermodynamic stability (the free energy of the secondary structure). Some SVM classifiers also use additional properties, such as entropy measures.

The web server for the BayesMiRNAfind program described in [30] also implements an SVM version of the classifier which follows the same pipeline as the Naïve Bayes classifier but uses the SVM instead of the Naïve Bayes model. As previously mentioned in Section 4.5.1, the BayesMiRNAfind feature vector uses only features based on sequence composition and topological properties.

One of the first successful miRNA detection programs to use the SVM classifier is described in the paper by Sewer et al. [20]. The method starts by moving a sliding window across the input RNA sequence finding the secondary structure with the minimal free energy for each window. The preservation rate, or robustness, for a nucleotide pair (i, j) is defined as the number of windows containing the nucleotide pair (i, j) divided by the number of windows containing both the nucleotides i and j. A minimal robustness value is chosen and used to filter out genomic regions that are not "robust" enough. The program then calculates the feature vector for each stem loop and classifies the stem loop using the SVM. The feature vector consists of four groups of features depending on which portion of the structure the statistic is computed over: the entire stem loop structure, the longest symmetrical region of the stem, the longest "relaxed symmetry region", or all of the windows on the candidate stem loop that correspond to the length of a mature miRNA. The "relaxed symmetry region" is defined as an asymmetrical loop region where the lengths of the two sides of the loop do not exceed a specified threshold. The 40 features used by Sewer et al. are listed in Tables 3-6. Figure 10 shows the SVM score distributions reported in their paper.

Table 3. Features calculated over entire stem loop structure [20]

1	Free energy of folding
2	Length of the longest simple stem
3	Length of the hairpin loop
4	Length of the longest perfect stem
5	Number of nucleotides in symmetrical loops
6	Number of nucleotides in asymmetrical loops
7	Average distance between internal loops
8	Average size of symmetrical loops
9	Average size of asymmetrical loops
10/11/12/13	Proportion of A/C/G/U nucleotides in the stem
14/15/16	Proportion of A-U/C-G/G-U base pairs in the stem

Table 4. Features calculated over longest symmetrical region [20]

17	Length
18	Distance from the hairpin loop
19	Number of nucleotides involved in internal loops
20/21/22/23	Proportion of A/C/G/U nucleotides
24/25/26	Proportion of A-U/C-G/G-U base pairs

Table 5. Features calculated over longest relaxed symmetry region [20]

27	Length
28	Distance from the hairpin loop
29	Number of nucleotides involved in symmetrical internal loops
30	Number of nucleotides involved in asymmetrical internal loops
31/32/33/34	Proportion of A/C/G/U nucleotides
35/36/37	Proportion of A-U/C-G/G-U base pairs

Table 6. Features calculated over all potential mature miRNA regions [20]

38	Maximum number of base pairs
39	Minimum number of nucleotides in asymmetrical loops
40	Minimum asymmetry over the internal loops in this region

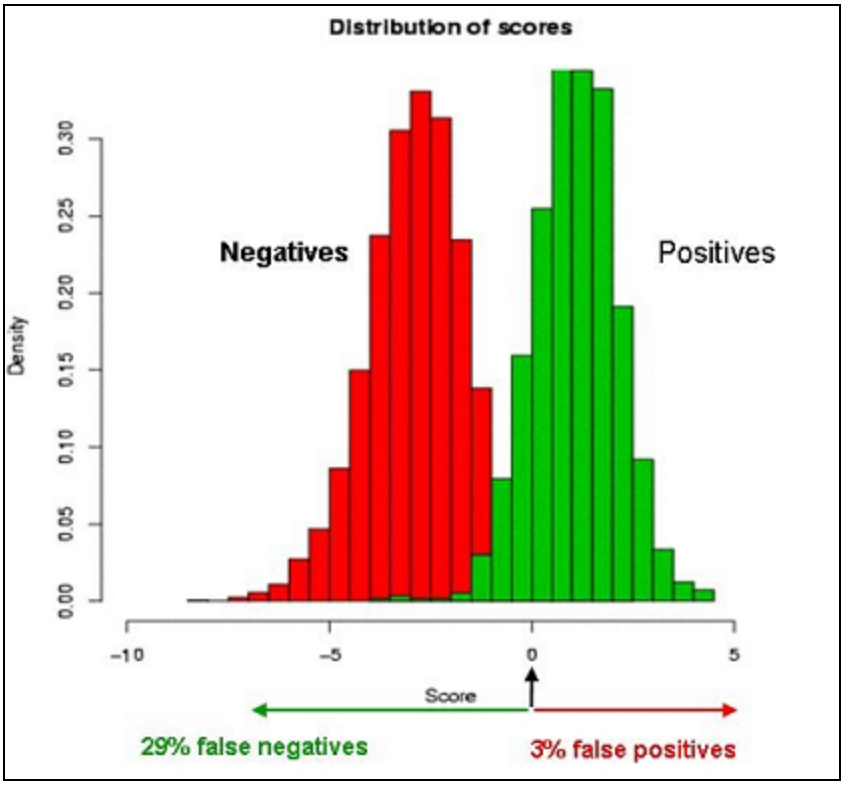

Figure 10. SVM score distributions [20]

The microPred tool developed by Batuwita & Palade [1] is an extension of the miPred developed by Ng & Mishra [18] (this is different from the MiPred tool that I use in my experiments). They used the 29 features from miPred and added 19 new features to the feature vector to try to improve the system's performance. The original feature vector of 29 "RNA global and intrinsic folding attributes" consists of 17 base composition variables, six folding measures, one topological descriptor and five normalized features obtained from performing dinucleotide shuffling [18]. The 19 new features introduced by microPred consist of two MFE-related features, four RNAfold-related features, six Mfold-related features and seven base pair related features [1].

4.5.4 Other Machine Learning Techniques

A miRNA predictor based on a novel machine learning technique, called random forests, is described in [12]. The random forest classification is the majority vote of a group of tree-structured classifiers that were trained on bootstrapped samples of the training data. The 34 features used by this algorithm are a minimum free energy (MFE) feature, a "P-value of randomization test feature", and 32 features representing "local contiguous triplet structure composition" [12]. Calculating the P-value relies on a process the authors called dinucleotide shuffling, which is where the order of the nucleotides in the sequences are randomized while keeping the dinucleotide frequencies constant. The P-value is defined as the ratio of shuffled sequences whose secondary structure has a lower MFE than the original sequence. Each nucleotide is either paired or unpaired in the sequence's secondary structure, represented as '(' and '.' respectively. So for each triplet of nucleotides, there are $2^3 = 8$ possibilities. There are 4 possible values for the middle nucleotide (A, C, U, G). For each of these 4 * 8 = 32 combinations, the number of times that that element occurs in the sequence makes up another feature in the feature vector. When growing a tree, only a subset of features is selected instead of using all of the features. The authors of this paper claim that the Random Forest classifier achieved 93.21% specificity and 89.35% sensitivity. However, one significant disadvantage of this technique is that calculating the P-value requires performing dinucleotide shuffling on the original sequence 1000 times, which is a very time consuming process.

Another novel method based on relaxed variable kernel density estimation (RVKDE) based classifiers, a special type of neural network, is described by Chang et al. [4]. The authors claim that the RVKDE classifier "exploits more local information of the training dataset" as compared to SVMs.

Another machine learning technique, linear genetic programming, has been used to automatically create and adapt special classifier programs that combine multiple structure motifs, each represented by a regular expression. The advantage of this method is that the motif can be scanned against the genome without having to pre-select potential stem-loop structures since matching a sequence to a motif is position-independent. The authors claimed that by using 16 motif-based classifiers, they could achieve 99.9% specificity with an acceptable high level of sensitivity, making it "at least competitive to state-of-the-art feature-based methods for ab initio miRNA discovery" [3].

5. Materials

5.1 Datasets

My full test dataset consisted of 1,442 human RNA sequences. My experiments required two datasets: a positive dataset and a negative dataset. Each dataset contained 721 sequences. No training dataset was required for my experiments since the web servers I was performing my experiments on were already trained. The positive dataset

was used to calculate the true positive (TP) and false negative (FN) statistics and the negative dataset was used to calculate the true negative (TN) and false positive (FP) statistics.

5.1.1 Positive Dataset

The positive dataset consisted of sequences of experimentally verified human miRNA precursors. For my project, I obtained the sequences for the positive dataset from miRBase [8]. Release 14 was released in September 2009 and contains over 10,000 entries in 115 species. miRBase contains 721 entries for human miRNAs and my positive dataset consisted of all 721 of those miRNA precursor sequences.

5.1.2 Negative Dataset

The negative dataset consisted of sequences of human pseudo pre-miRNA. These represent sequences that have similar properties to actual pre-miRNA sequences such as the stem-loop secondary structure but are not known to be actual pre-miRNAs. The sequences I used are a fraction of the dataset generated by Xue et al. [27]. These sequences were extracted from the protein coding sequences (CDS) from human genes and the full dataset contained 8,494 of these pseudo pre-miRNA sequences. For my experiments, I simply took the first 721 of these sequences to make the size of the positive and negative datasets equal.

5.2 Software

5.2.1 BayesSVMmiRNAfind

The BayesSVMmiRNAfind web server runs both the Naïve Bayes and SVM based classifiers for the system described Yousef et al. [30]. The web address is http://wotan.wistar.upenn.edu/BayesSVMmiRNAfind/. Figure 11 shows the input screen for the web server and Figure 12 shows a sample predicted miRNA.

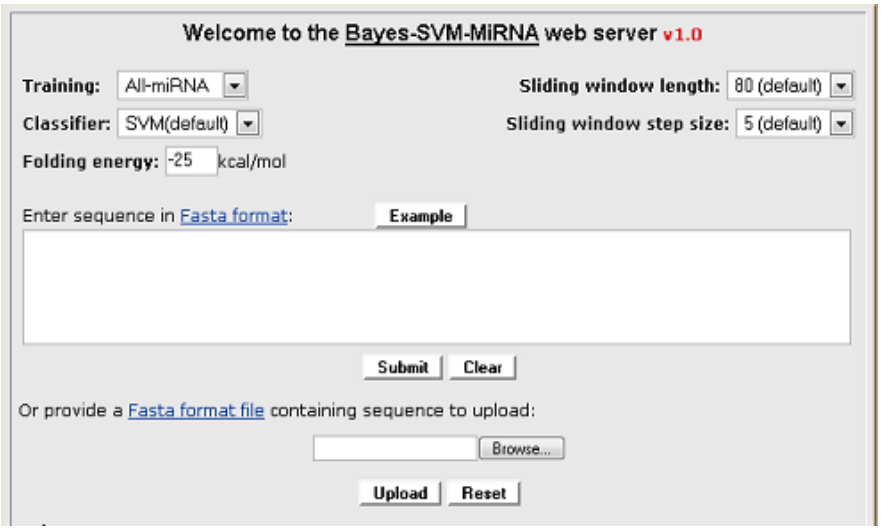

Figure 11. BayesSVMmiRNAfind input screen

hsa-let-7a-2___SEQUENCE_2 prediction found

Prediction 1	hsa-let-7a-2___SEQUENCE_2
Name	Value
SVM Score:	1.8926
Sliding window:	0-73
Sequence length:	72
Folding energy:	-25.1
Loop length:	6
Precursor sequence:	aggttgaggtagtaggttgtatagtttagaattacatcaagggagataactgtacagcctcctagctttcct
miRNA:	aactgtacagcctcctagctt (Location: 47-67)
Secondary structure: (w/o loop)	5' aggttgaggtagtaggttgtatagtttagaattacatc 3' 3' tcct-ttcgatcctccgacatgtcaa---------tag 5'

Figure 12. Sample BayesSVMmiRNAfind prediction

5.2.2 miR-abela

The SVM based classifier described by Sewer et al. [20] is run on the miR-abela web server (http://www.mirz.unibas.ch/cgi/pred_miRNA_genes.cgi). Figure 13 shows the input screen for the web server and Figure 14 shows some sample predicted miRNAs. The web server also allows the user to enter an email address for large batch sequences and the results will be emailed to the user.

Figure 13. miR-abela input screen

Name	Genomic Loci	Score	Length	Energy	Sequence and secondary structure
hsa-mir-7-2___SEQUENCE_15-P1	chr|+|442|539	2.42	98	-43.10	GUGGACCGGCUGGCCCCAUCUGGAAGACUAGUGAUUUUGUUGUUGUCUUACUGCGCUCAACAACAAAUCC
hsa-mir-10a___SEQUENCE_20-P1	chr|+|435|531	2.04	97	-32.29	UGUCUGUCUUCUGUAUAUACCCUGUAGAUCCGAAUUUGUGUAAGGAAUUUUGUGGUCACAAAUUCGUAUC
hsa-mir-16-1___SEQUENCE_24-P1	chr|+|431|519	1.91	89	-37.70	GUCAGCAGUGCCUUAGCAGCACGUAAAUAUUGGCGUUAAGAUUCUAAAAUUAUCUCCAGUAUUAAC
hsa-mir-9-3___SEQUENCE_19-P1	chr|+|432|519	1.77	88	-39.40	GAGGCCCGUUUCUCUCUUUGGUUAUCUAGCUGUAUGAGUGCCACAGAGCCGUCAUAAAGCUAGAU
hsa-mir-1-2___SEQUENCE_13-P1	chr|+|432|514	1.68	83	-34.74	CCUACUCAGAGUACAUACUUCUUUAUGUACCCAUAUGAACAUACAAUGCUAUGGAAUGUAAAC
hsa-mir-7-1___SEQUENCE_14-P1	chr|+|437|532	1.62	96	-44.40	GUUGGCCUAGUUCUGUGUGGAAGACUAGUGAUUUUGUUGUUUUUAGAUAACUAAAUCGACAACAAAUCA

Figure 14. Sample miR-abela predictions

5.2.3 MiPred

The MiPred web server (http://www.bioinf.seu.edu.cn/miRNA/) runs the random forest based classifier described by Jiang et al. [12]. Unfortunately, due to significant limitations on that particular web server (only being able to run 3 sequences at a time) and very long computation times, I was only able to test a small fraction of my sequences with MiPred. Figure 15 shows the input screen for the web server and Figure 16 shows a sample predicted miRNA.

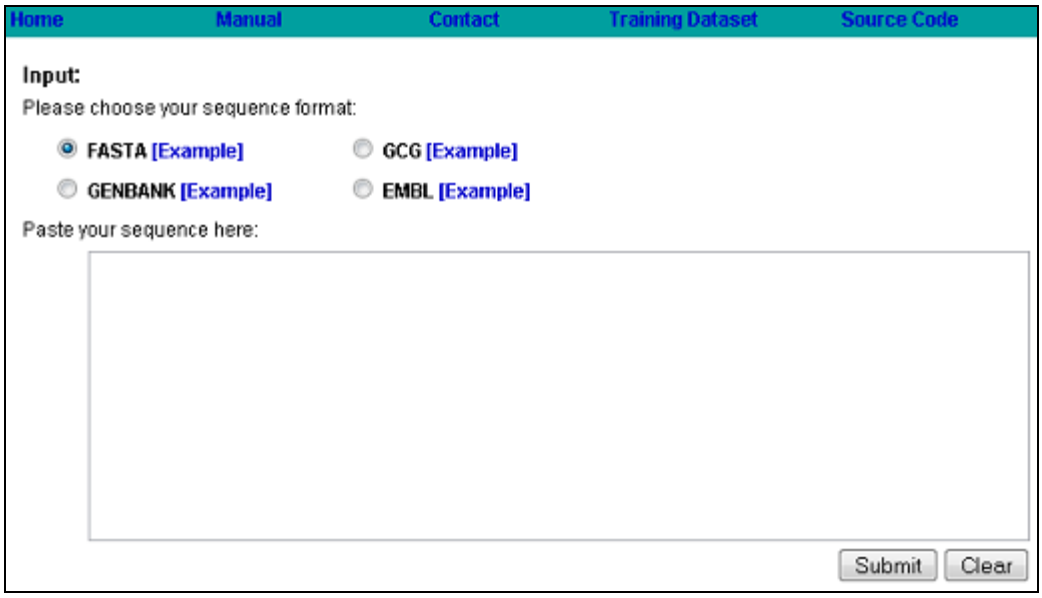

Figure 15. MiPred input screen

Sequence Name:	hsa-mir-24-1___SEQUENCE_38
Sequence Content:	CUCCGGUGCCUACUGAGCUGAUAUCAGUUCUCAUUUUACACACUGGCUCAGUUCAGCAGGAACAGGAG
Length:	68
Pre-miRNA-like Hairpin?	Yes
The Secondary Structure:	((((.((.(((.(((((((((..(((((.....))))).)))))))))).))).)).))))
MFE:	-26.32
p-value (shuffle times:1000)	0.001
Prediction result:	Real microRNA precursor
Prediction confidence:	73.9%

Figure 16. Sample MiPred prediction

5.2.4 microPred

The microPred web server runs the SVM classifier described in [1] and simply takes an email address and a text file containing up to a hundred RNA sequences in FASTA format and emails the results when it's finished. The web address is http://www.comlab.ox.ac.uk/microPred/microPred-server.html.

6. Analysis

Five statistics were calculated from the results of my experiments to measure and compare the performance of each tool. Sensitivity (Se) measures the number of actual positives that were predicted as being positive. Specificity (Sp) measures the number of actual negatives that were predicted as being negative. Accuracy (Acc) measures the proportion of correct predictions. The Matthews Correlation Coefficient (MCC) measures the quality of binary classifications. The Positive Predictive Value (PPV) measures the proportion of positive predictions that were correctly predicted.

Acc, Se, Sp, and PPV all return percentage values between 0 and 100 where higher numbers represent more accurate predictions. MCC is commonly used in machine learning and is considered to be a balanced measure and one of the most useful measures of a binary classifier. MCC returns a value between -1 and 1. "A coefficient of +1 represents a perfect prediction, 0 an average random prediction, and -1 an inverse

prediction" [26]. PPV is also a useful tool for analyzing miRNA prediction tools because experimental verification of these predictions can be very difficult and time consuming but if the tool has a very high PPV then the user can have more confidence in the prediction. The equations for the five performance evaluators, which were taken from Sinha et al. [22], are shown below:

TP = True Positive; FP = False Positive; TN = True Negative; FN = False Negative

$$Se = \frac{TP}{(TP + FN)} * 100$$

$$Sp = \frac{TN}{(TN + FP)} * 100$$

$$Acc = \frac{(TP + TN)}{(TP + TN + FP + FN)} * 100$$

$$MCC = \frac{(TP * TN) - (FP * FN)}{\sqrt{(TP + FP) * (TN + FN) * (TP + FN) * (TN + FP)}}$$

$$PPV = \frac{TP}{(TP + FP)} * 100$$

7. Results

Table 7 shows the TP, FN, TN and FP values from my experiments with the miRNA prediction tools. Table 8 shows the values for the performance evaluation indicators: accuracy, sensitivity, specificity, MCC and PPV. Due to long computation times, not all tools were run on the entire dataset. MiPred was only tested on the first 75

of 721 sequences from both the positive and negative datasets (10.4% of the entire dataset). microPred only has results for 600 of 721 sequences in the positive dataset (83.2%) and 500 of 721 sequences in the negative dataset (69.3%).

Table 7. TP, FN, TN and FP values for miRNA predictions

Tools	Positive Dataset		Negative Dataset	
	TP	FN	TN	FP
1. BayesSVMmiRNAfind				
Naïve Bayes	97.36	2.64	28.99	71.01
SVM	99.31	0.69	14.29	85.71
2. miR-abela	62.55	37.45	99.86	0.14
3. MiPred	93.33	6.67	94.67	5.33
4. microPred	89.17	10.83	74.80	25.20

Table 8. Performance of miRNA detection tools

Tools	Se	Sp	Acc	MCC	PPV
1. BayesSVMmiRNAfind					
Naïve Bayes	97.36	28.99	63.18	0.3611	57.83
SVM	99.31	14.29	56.80	0.2582	53.67
2. miR-abela	62.55	99.86	81.21	0.6727	99.78
3. MiPred	93.33	94.67	94.00	0.8801	94.59
4. microPred	89.17	74.80	82.64	0.6504	80.94

BayesSVMmiRNAfind has the highest sensitivity but also has the lowest specificity, accuracy, MCC and PPV. miR-abela has the lowest sensitivity but it also has the highest specificity and PPV. MiPred showed high sensitivity and specificity and also has the highest accuracy and MCC. microPred showed higher sensitivity but lower specificity than miR-abela but both tools had similar accuracy and MCC with miR-abela having the higher PPV of the two tools.

8. Discussion

Although BayesSVMmiRNAfind had the highest sensitivity, it had very low specificity which significantly lowered its MCC and PPV values. With such a low PPV, 57.83% for Naïve Bayes and 53.67% for SVM, the user should not have a lot of confidence in any novel miRNAs predicted by BayesSVMmiRNAfind, making it a less valuable tool compared to the other three programs. MiPred was the only one of the four tools to achieve a very high score in both sensitivity and specificity (93.33% and 94.67% respectively). Its MCC value of 0.8801 was the highest of the four tools and surpassed the second highest MCC score by over 30%. MiPred also achieved the second highest PPV value at 94.59%. miR-abela and microPred achieved comparable accuracy and MCC values. Although miR-abela had the highest PPV (99.78%), it low sensitivity (62.55%) means that it would not be very effective for detecting novel miRNAs since it would end up missing too many actual miRNAs in its predictions. microPred has a PPV of 80.94% but because it has 89.17% sensitivity, it would probably be more likely to find novel miRNAs compared to miR-abela although it would also pick up more false positives.

Table 9 shows a summary of the four miRNA detection tools that were tested. As previously mentioned, both MiPred and microPred perform the time-consuming process of dinucleotide shuffling which makes them take much longer to analyze an RNA sequence compared to the other two tools.

Table 9. Summary of miRNA detection tools used

Classifier	Total # of Features	Sequence Composition	Topological Properties	Thermodynamic Stability	Entropy Measures
BayesSVM [30]	74	X	X		
miR-abela [20]	40	X	X	X	
MiPred [12]	34	X	X	X	
microPred [1]	48	X	X	X	X

These results show that the quality of the features is more important than just the total number of features. BayesSVMmiRNAfind had the most features of the four programs but had the worst performance while the program with the best MCC score, MiPred, had the least amount of features.

These results also show that adding the correct features to the feature vector of an existing system can provide very good results. The 32 structure-sequence features used by MiPred were used as the feature vector for a system called triple-SVM, which is discussed in [29]. When the MFE and P-value features were added to the feature vector, it significantly improved both the sensitivity and specificity, as reported in [12]. Switching from the SVM classifier to the Random Forest model provided further improvements to the tool's performance. As previously discussed in this report, microPred also expanded the feature vector of an existing tool to improve that tool. The authors of microPred reported that the new feature vector improved the sensitivity by nearly 9% while leaving the specificity at the same level as the original feature vector [1].

A final conclusion that I could draw from the results of my experiments is that dinucleotide shuffling is a very powerful process even though it is very computationally intensive. The two tools that used dinucleotide shuffling, MiPred and microPred, were the only two tools to achieve a high sensitivity value while also maintaining an acceptably high specificity. Jiang et al. reported that the P-value feature was the most important feature for distinguishing between real and pseudo pre-miRNA hairpins. They claim that "random RNA must be generated with the same dinucleotide frequency for any valid conclusion to be drawn" so the purpose of the P-value feature is "to determine if the MFE value is significantly different from that of random sequences" [12]. This is also supported by Ng and Mishra who reported that two of the top four most important features were normalized features [18]. Although, dinucleotide shuffling is very time consuming, the fact that each individual shuffle could be performed independently, as long as the results are aggregated properly, makes this step ideal for multi-threaded computing. Being able to split the computation on a dual-core machine alone could potentially cut the runtime nearly in half.

9. Conclusion

This work has provided the motivation behind the development of computation methods for detecting miRNA genes and presented an overview of many different methods that have been developed. Homology-based searches are only capable of detecting miRNAs that are homologues of known miRNAs. Filtering-based methods rely

on sequence and structure conservation and are limited by a lack of a clear evolutionary model. Target-centered approaches rely on highly conserved motifs in the 3'-UTRs but make few assumptions about the pre-miRNA structure. Many different machine learning methods have been applied to miRNA gene prediction. These methods have been shown to be able to achieve good scores for both sensitivity and specificity and their performance is expected to improve as new miRNAs are identified and added to their training data. This makes machine learning based miRNA detectors a very important tool for detecting novel miRNAs. Of the four tools that I analyzed, MiPred achieved the best performance with the highest accuracy and MCC values. BayesSVMmiRNAfind achieved the highest sensitivity but also the lower specificity values. miR-abela achieved the highest specificity and PPV but its relatively low sensitivity decreases the tool's usefulness. microPred achieved more balanced sensitivity and specificity compared to miR-abela. My experiments showed that having a feature vector with good features is more important than just padding the feature vector with less important features. Although dinucleotide shuffling was shown to be very important in improving the performance of miRNA detection tools, it has very long computation times. Fortunately, the process seems to be well-suited for multi-threaded computing and could benefit significantly from distributing the workload, which would help to close the gap between the runtime of tools with and without dinucleotide shuffling. Additional future work could involve investigating the usefulness of chaining the results from a faster but less accurate tool to a slower tool with better accuracy.

References

1. Batuwita, R. & Palade, V. (2009) microPred: Effective Classification of pre-miRNAs for Human miRNA Gene Prediction. *Bioinformatics*, **25** (8): 989-995.

2. Berezikov, E., Guryev, V., van deBelt, J., Wienholds, E., Plasterk, R.H.A., & Cuppen, E. (2005) Phylogenetic Shadowing and Computation Identification of Human MicroRNA Genes. *Cell*, **120** (1): 21-24.

3. Brameier, M. & Wiuf, C. (2007) Ab Initio Identification of Human MicroRNAs Based on Structure Motifs. *BMC Bioinformatics*, **8**: 478.

4. Chang, D., Wang, C.-C. & Chen, J.-W. (2008) Using a Kernel Density Estimation Based Classifier to Predict Species-Specific MicroRNA Precursors. *BMC Bioinformatics*, **9** (Suppl 12): S2.

5. Chaudhuri, K., & Chatterjee, R. (2007) MicroRNA Detection and Target Prediction: Integration of Computational and Experimental Approaches. *DNA and Cell Biology*, **26** (5): 321-337.

6. Gautheret, D. & Lambert, A. (2001) Direct RNA Motif Definition and Identification from Multiple Sequence Alignments Using Secondary Structure Profiles. *Journal of Molecular Biology*, **313** (5): 1003-1011.

7. Griffiths-Jones, S., Moxon, S., Marshall, M., Khanna, A., Eddy, S.R., & Bateman, A. (2005) Rfam: Annotating Non-Coding RNAs in Complete Genomes. *Nucleic Acids Research*, **33** (Database issue): D121-D124.

8. Griffiths-Jones, S., Saini, H.K., van Dongen, S. & Enright, A.J. (2008) miRBase: Tools for MicroRNA Genomics. *Nucleic Acids Research*, **36** (Database Issue): D154-D68.

9. Helvik, S.A., Snøve, O. & Saetrom, P. (2007) Reliable Prediction of Drosha Processing Sites Improves MicroRNA Gene Prediction. *Bioinformatics*, **23** (2): 142-149.

10. Hertel, J., & Stadler, P.F. (2006) Hairpins in a Haystack: Recognizing MicroRNA Precursors in Comparative Genomics Data. *Bioinformatics*, **22** (14): e197-e202.

11. Huang, T., Fan, B., Rothschild, M., Hu, Z., Li, K. & Zhao, S. (2007) MiRFinder: an Improved Approach and Software Implementation for Genome-Wide Fast MicroRNA Precursor Scans. *BMC Bioinformatics*, **8**: 341.

12. 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-W344.

13. Lai, E.C. (2003) MicroRNAs: Runts of the Genome Assert Themselves. *Current Biology*, **13** (23): R925-R936.

14. Legendre, M., Lambert, A., & Gautheret, D. (2005) Profile-Based Detection of MicroRNA Precursors in Animal Genomes. *Bioinformatics*, **21** (7): 841-845.

15. Mendes, N.D., Freitas, A.T. & Sagot, M.-F. (2009) Current Tools for the Identification of miRNA Genes and their Targets. *Nucleic Acids Research*, **37** (8): 2419-2433.

16. Nam, J., Shin, K., Han, J., Lee, Y., Kim, V. & Zhang, B. (2005) Human MicroRNA Prediction through a Probabilistic Co-Learning Model of Sequence and Structure. *Nucleic Acids Research*, **33** (11): 3570-3581.

17. Nam, J., Kim, J., Kim, S., & Zhang, B. (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-W458.

18. 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*, **23** (11): 1321-1330.

19. Setubal, J. & Meidanis, J. (1997) *Introduction to Computational Molecular Biology*. PWS Publishing Company.

20. Sewer, A., Paul, N., Landgraf, P., Aravin, A., Pfeffer, S., Brownstein, M.J., Tuschl, T., vanNimwegen, E. & Zabolan, M. (2005) Identification of Clustered MicroRNAs Using an Ab Initio Prediction Method. *BMC Bioinformatics*, **6**: 267.

21. Sheng, Y., Engström, P.G. & Lenhard, B. (2007) Mammalian MicroRNA Prediction through a Support Vector Machine Model of Sequence and Structure. *PLoS ONE*, **2** (9): e946.

22. Sinha, S., Vasulu, T.S. & De, R.K. (2009) Performance and Evaluation of MicroRNA Gene Identification Tools. *Journal of Proteomics and Bioinformatics*, **2** (8): 336-343.

23. Terai, G., Komori, T., Asai, K., & Kin, T. (2007) miRRim: A Novel System to Find Conserved miRNAs with High Sensitivity and Specificity. *RNA,* **13** (12): 2081-2090.

24. Wang, X., Zhang, J., Li, F., Gu, J., He, T., Zhang, X. & Li, Y. (2005) MicroRNA Identification Based on Sequence and Structure Alignment. *Bioinformatics*, **21** (18): 3610-3614.

25. Wikipedia Contributors. "MicroRNA." *Wikipedia, The Free Encyclopedia*, http://en.wikipedia.org/wiki/MicroRNA (accessed May 17, 2009).

26. Wikipedia Contributors. "Matthews correlation coefficient." *Wikipedia, The Free Encyclopedia*, http://en.wikipedia.org/wiki/Matthews_Correlation_Coefficient (accessed November 21, 2009).

27. Wong, K. M. (2009) New Respect for mRNA. *ScienceMatters@Berkeley*, **6** (41).

28. Xie, X., Lu, J., Kulbokas, E.J., Golub, T.R., Mootha, V., Lindblad-Toh, K., Lander, E.S. & Kellis, M. (2005) Systematic Discovery of Regulatory Motifs in Human Promoters and 3'UTRs by Comparison of Several Mammals. *Nature*, **434**: 338-345.

29. Xue, C., Li, F., He, T., Liu, G., Li, Y. & Zhang, X. (2005) Classification of Real and Pseudo MicroRNA Precursors Using Local Structure-Sequence Features and Support Vector Machines. *BMC Bioinformatics*, **6**: 310.

30. Yousef, M., Nebozhyn, M., Shatkay, H., Kanterakis, S., Showe, L.C. & Showe, M.K. (2006) Combining Multi-Species Genomic Data for MicroRNA Identification Using a Naïve Bayes Classifier. *Bioinformatics*, **22** (11): 1325-1334.

31. Zvelebil, M. & Baum, J.O. (2008) *Understanding Bioinformatics*. Garland Science.