Increasing the Target Prediction Accuracy of MicroRNA Based on Combination of Prediction Algorithms

Mohammed Q. Shatnawi*
Computer Information Systems Dept.
Jordan University of Science and Technology
Irbid, Jordan

Mohammad Alhammouri, Kholoud Mukdadi
Computer Science Department
Jordan University of Science and Technology
Irbid, Jordan

***Abstract*—MicroRNA is an oligonucleotide that plays a role in the pathogenesis of several diseases (mentioning Cancer). It is a non-coding RNA that is involved in the control of gene expression through the binding and inhibition of mRNA.**

In this study, three algorithms were implemented in WEKA software using two testing modes to analyze five datasets of miRNA families. The data mining techniques are used to compare the interactions of miRNA-mRNA that it either belongs to the same gene-family or to different families, and to establish a biological scheme that explains how the biological parameters are involved or less involved in miRNA-mRNA prediction.

The factors that were involved in the prediction process includs match, mismatch, bulge, loop, and score to represent the binding characteristics, while the position, 3'UTR length, and chromosomal location and chromosomal categorizations represent the characteristics of the target mRNA. These attributes can provide an empirical guidance for study of specific miRNA family to scan the whole human genome for novel targets. This research provides promising results that can be utilized for current and future research in this field.

Keywords—miRNA; chromosome; prediction; genome; disease; biology; DNA sequence; enzyme

I. Introduction

The cell, the basic unit of a living organism, has an extraordinary ability to reproduce, grow, respond to stimuli, and exchange a wide range of materials (metabolites) with its surrounding environment. All these tasks have been found to be heritable in nature from parent cells to progenies, and have been explained through a molecular model that has become the dogma of molecular biology.

The majority of the functions inside the cells are carried out by enzymes (Proteins and sometimes RNA). These components are orchestrated and controlled by genes. In simple terms, genes are heritable codes in the form of DNA sequence that can be transcribed into mRNA and then translated into proteins. This process is described as gene expression. Gene expression determines the fate of a cell because it controls the types of RNA and Proteins and also their exact amounts in a certain cell. Thus, all the cells in a human being contain the same DNA (genes) but due to differences in gene expression, a cell could become either a heart or a brain, or a muscle cell (with few exceptions).

Finally, Proteins is composed of amino acids, and are constructed by ribosome according to the genetic code in a process called translation. Each of the 20 amino acids is represented in the genetic code as three nitrogen bases. Proteins are the functional component of the cell; they play a role in building the structure of the cell, as enzymes (catalyzing chemical reactions), as signaling molecules (hormones), or transport vehicles, etc [1]. Messenger RNA transcript is composed of several components as discussed in [1], also miRNA can be defined as a class of short non-coding RNAs that are approximately 21 nucleotides (nt) in length [2].

A. Regulation of Gene Expression

Gene expression is regulated through a wide range of factors along with the different steps of transcription and translation. The first step of regulation occurs at the chromatin level. Chromatin is the natural packaging of DNA with special proteins that protects and controls genes inside the cell through several chemical modifications of the chromatin Histone proteins (e.g. Histone acetylating activates genes while Methylation represses them etc.) and DNA itself. DNA Methylation causes gene silencing.

Another level of control occurs at the DNA sequence level though the promoter and other upstream regulatory elements which act as a binding site ,that is (called cis-elements) for transcription factors (trans-elements) which are complexes of proteins that recruit or repress the enzymes of transcription (i.e. RNA polymerases).

After transcription, mRNA levels are controlled by its half-life through several factors. (miRNA) is believed to be a key player in the control of mRNA levels. miRNA binds to mRNA in a specific fashion and recruits a number of enzymes responsible for mRNA degradation.

Finally, which expression can be regulated at the protein level either by chemical modifications of the proteins (eg. phosphorylation) or by a feedback loop such as in the case of transcription factors that control their own genes. [1]

B. Functions of miRNA

(miRNA) plays a great role in Arranging a large number of target genes in the gene expression. Most of human miRNA genes have been defined and combined to realize a stable number of functions.

Let-7 and Lin4 miRNA are considered as some of the first discovered miRNA genes in the C.elegans worm. Talking about human and other vertebrate cell lines, tumor suppression, antiviral defense, adiposity differentiation and susceptibility to cytotoxic T-cells include some miRNA genes [2].

The MicroRNAs play important and critical roles in genetic human diseases [3, 4]. Such as, breast cancer [5] and heart diseases [6]. Nevertheless, the mechanism of gene expression that is regulated by miRNA remains ambiguous.

C. Problem Statement

Predicting miRNA-mRNA interaction by experiment is highly costly in terms of time, labor and mone. The number of miRNA and mRNA studies is increasing on a daily basis. Attempts to design biological experiments are always rendered outdated once the experiment is about to start. In addition, experiments may lack the element of discovering the novel miRNA-mRNA interactions. Thereby, data that is stacking in biological databases should be invested to discover new miRNA-mRNA interactions. The problems in miRNA-mRNA target prediction can be summarized in the following points:

- Previous studies in the field of miRNA–mRNA target prediction are scattered in different aspects of the binding process such as sequence complementarily, the binding energy (thermodynamics), etc. This limits the involvement of the factors that play a role in the prediction.
- Previous studies often use one classification method, which gives great weight to specific features on the account of others.
- miRNA are uniquely involved in complex cellular pathways that include well organized and controlled networks of genes. These networks may or may not “cross-talk” with each other. This fact was not taken into consideration in the previous studies and miRNAs were studied in bulk and not as separate families.

D. Problem Solution

Several attempts were used to apply bioinformatics techniques in order to find an optimized algorithm that can be used to explain miRNA-mRNA interaction mathematically as well as biologically. At mean time, there are several studies attempting to optimize miRNA-mRNA prediction from different aspects ranging from RNA sequence to binding energy (thermodynamics). Such algorithms can be used to screen the genome for new mRNA targets and predict miRNA-mRNA relationships that can be of a great benefit to the treatment and diagnosis of many diseases. In this study, the following solutions are taken into consideration for the prediction of miRNA-mRNA relationships:

- Factors from different aspects of miRNA-mRNA binding are given in this study, such as sequence complementarily, the binding energy (thermodynamics), etc..
- Three classification methods (i.e. decision tree, naïve bayes and support vector machine) are used.
- In order to shed the light on the uniqueness of miRNAs, five families of miRNA are studied as one collection and each one as a standalone family.

E. Knowledge discovery in Database Process

Knowledge discovery in database (i.e. KDD) is a very important process, which is the general process of converting raw data into useful information. The data mining is an integral part of this process [7].

Therefore, there are enormous and massive collection of data that is stored about the genes, proteins, and other vital information for each human being. As a result, KDD process can be applied to extract information, patterns, and new rules using different techniques [8].

F. Motivation for miRNA Research

Recently, miRNA has been found to play a key role in most cellular pathways. Thus, it is now considered one of the basic tools of gene expression regulation. miRNA has changed our textbook view of the biological process until it finally forced itself on the frontier of the biomedical sciences. One of the motivations for establishing this study is the existing of Princess Haya Biotechnology Centre (PHBC); is which a well equipped research center for biological study of miRNA where the results of this work can be experimentally validated and applied. This study can contribute to different fields of biology and medicine such as, the study of disease pathogenesis, animal models, targeted drag, drug treatment and relationship between cytogenetic (study of chromosomes) and epigenetic (study of heritable changes inflected by factors like miRNA).

G. Objectives

The main goal of this study is to use data mining for predicting the miRNA-mRNA interaction through the implementation of the following objectives:

1) Collecting the miRNA and mRNA data from databases. These data include biological parameters that are related to sequence, chromosomal location, structure folding and previous known interaction scores.

2) The use of different data mining techniques to study and compare the interaction of miRNA that is either belonging to the same gene-family or to different families.

3) Establish a biological scheme that can explain how the biological parameters are involved or less involved in miRNA-mRNA prediction.

H. Significant Contributions

This study provides an insight to the biological parameters that are involved or neglected in miRNA-mRNA target prediction, and shed some light on the mechanisms that are underlying gene silencing in cancer cellular pathways.

This study is rather significant from a clinical perspective; the establishment of a good miRNA-mRNA prediction tool can help in discovering novel gene interactions, which can open the gate for new drug targets, and novel mutated disease genes. Thus, pushing forward the process of disease treatment and diagnosis is in progress.

On the other hand, this field is still in its infancy and the nature of miRNA-mRNA interaction is still not yet understood. The establishing of new parameters that are involved in this interaction gives more light attention to the biology behind the process of gene regulation.

II. Related Work

In literature, there are several researches that have been done on the miRNAs to predict their putative target mRNAs. They have been classified into different categories: researches based on computational method and probabilistic models, and researches based on machine learning methods.

A. Researches Based on Computational Method and Probabilistic Models

Hasan Og˘ul, et al. in [9] introduced a probabilistic model to show the binding preferences of miRNA and its predicted target. This model transforms an aligned duplex to represent a new sequence and used a Variable Length Markov Chain (VLMC) to determine the possibility of this sequence.

In [10], Chenghai Xue et al. Proposed a computational method to find the functional miRNA–mRNA regulatory modules (FMRMs) and to collect the miRNA in normal case and prostate cancer as a case to study the method contains groups of miRNAs and their putative target mRNAs under specific conditions. This computational method has successfully identified down-regulated patterns of mRNAs targets that are associated with prostate cancer and mRNAs associated with normal cases. Briefly, after preparing the dataset, authors applied association's algorithms in data mining to identify the biologically related miRNA–mRNA groups.

Wan Hsieh and Hsiuying Wang in [11] selected the human miRNA target prediction and suggested a generalized relative R2 method (RRSM) to discover many high-confidence prediction targets. RRSM is created based on relative rather than an absolute statistical view. In addition, it provides an efficient approach for miRNA target determination. RRSM program is available online at NCTU State website. [12]

In [13], William Ritchie et al. proposed an approach for the determination of putative miRNA targets based on a comparison between expression data of miRNAs and that of mRNAs using luciferase reporter assay. The miRNAs can decrease the expression level of targeted genes with direct correlation or indirect correlation between them. The success of this model was limited because the expression scalability of miRNA and mRNA was large. In addition, there are indirect functional relationships between two molecules.

Xiaofeng Song et al. in [14] proposed a computational method that is called microDoR to identify the mechanism of gene silencing by miRNA in humans, after they analyzed many features to find which are correlated with gene inhibition by miRNA. They found that the duplex structure of miRNA, the structural accessibility of mRNA target site region, and the numbers of binding sites are more efficient factors in identifying the target mRNA. The model that is based on SVM classifier is used to predict miRNA regulation based on these useful features. This study use all duplexes predicted by PicTar for a miRNA–mRNA interaction. The proposed approach in [14] was successful in distinguishing the mechanism by which the target mRNA is silenced either by cleavage or during translation.

In [15], Scott Younger et al. used computational methods for predicting possible miRNA targets through gene promoters and showed those promoters. Although, they are not conventionally linked to miRNA, they are strong candidates for miRNA regulation. Promoters are part of the non-transcribed sequence, and play a role in gene regulation prior to transcription. It is possible that functional correlation between promoter sequence and miRNA leads to a correlation between these sequences with their cellular pathways. Their study depends on seed sequence alignment. After that, they calculated the free energy scoring of miRNA and complementary scoring between miRNA and target sequence.

In [16], Alain Sewer et al. used a computational method to develop a program to approximate the pre-miRNA content and to predict the site of precursor miRNAs in genomic sequences. This program can be used to direct experiments to find both miRNAs that are evolutionarily conserved (i.e. miRNA has not been subjected to dramatic sequence changes because of the functional importance of the sequence) or added to species-specific miRNAs (i.e. miRNA was subjected to continuou evolutionary cycles of duplication and diversion leading to formation of sequences unique to specific species).

B. Researches Based on Machine Learning Methods

In [17], Xingqi Yan et al. used an ensemble machine learning algorithm in an attempt to improve prediction of miRNA targets. They used dataset from miRanda website [18].

This work has two major steps; in the first one, they input the biologically validated dataset, then they used feature selection (FS) to select the most informative features to be used as training data for the machine learning classifier. In addition, they used Adaptive Boosting (Adaboost) algorithm to create the ensemble classifiers that consist of several SVM classifiers to improve performance. In the second step, they used the previous one to apply this classifier on the result of miRanda. At each step, they performed evaluation processes.

In [19], Malik Yousef et al. described a target prediction method (NBmiRTar) using machine learning by a naïve Bayes classifier. The model is generated from sequence and miRNA:mRNA duplex information. Authors, used both the 'seed' and the 'out-seed' groups of the miRNA:mRNA duplex.

Their technique decreases false positive predictions and reduces the target possible number to be tested. In addition, the technique increases the sensitivity and specificity rather than the algorithms that depends on conserved genomic regions. The NBmiRTar software is available on NBmiRTar website [20]

Sung-Kyu Kim et al. created in [21] a miTarget model using a support vector machine (SVM) classifier for miRNA target gene prediction. miTarget depends on three categorized features. Those are structural, thermodynamic, and position-based features, which express the method of miRNA binding that were introduced in this study for the first time. This model produces high performance where it is compared with the previous tools. Authors are selected miR-1, miR- 124a, and miR-373 for humans using Gene Ontology (GO) analysis and discovered that significance of pairing at four, five and six positions in the 5' region of a miRNA are more essential than other seed regions.

In [22], Chenghai Xue et al. applied support vector machine to classify real versus pseudo pre-miRNAs depending on extracted features from hairpin local structure-sequence. The most important feature that defines miRNA precursors is the hairpin structures, but there are many similar hairpins that can be formed from genomic sequence. This classification helps to create a new approach to discover new miRNAs. In [33], the authors use the decision tree method to improve the accuracy of prediction. The authors discover the relationship between the data where the classified data are extracted based on rules.

It was still ambiguous when factors identify target silencing either due to transcript degradation or due to translational repression. Therefore, [23] defined two categories of target genes as the mRNA degradation (M-D) class and category the translational repression (T-R) class.

The authors of [24] noticed that the previous techniques of miRNA target prediction which were based only on sequence comparison resulted in many false positive interactions. In their work, they used the network context for filtering and support vector machine (SVM) because the topological characteristics of proteins in PPI (protein- protein interaction) networks can be used as another source of information for filtering out false positive miRNA target predictions.

III. Research Methodology

A. Overview

Most of the previous miRNA studies focused on the target prediction of miRNA binding using many features such as data mining or statistical techniques; trying to help and guide the experiments in the laboratory. Therefore, this study focuses on finding a correlation between the miRNA target sites of specific types of miRNA, namely; let-7a, let-7b, let-7c? family, mir-21 and mir-122 and the chromosomal location [i.e. q: long arm of chromosome and p : short arm of chromosome], the nucleotide sequence, binding and thermodynamic features of miRNA and mRNA. The dataset was collected from miRNA target prediction database (i.e. database using miRANDA algorithm) and then three techniques of data mining are used to investigate the miRNA-mRNA relationship by weka 3.6 software as shown in Fig.1.

All miRNA types in this study had the same length (22 nt) and were previously shown to play role in causing cancer. For instance mir-21 is involved in breast cancer pathogenesis.

In order to lower the false positives in this study, only the highest scored miRNA binding site (out of three) is included.

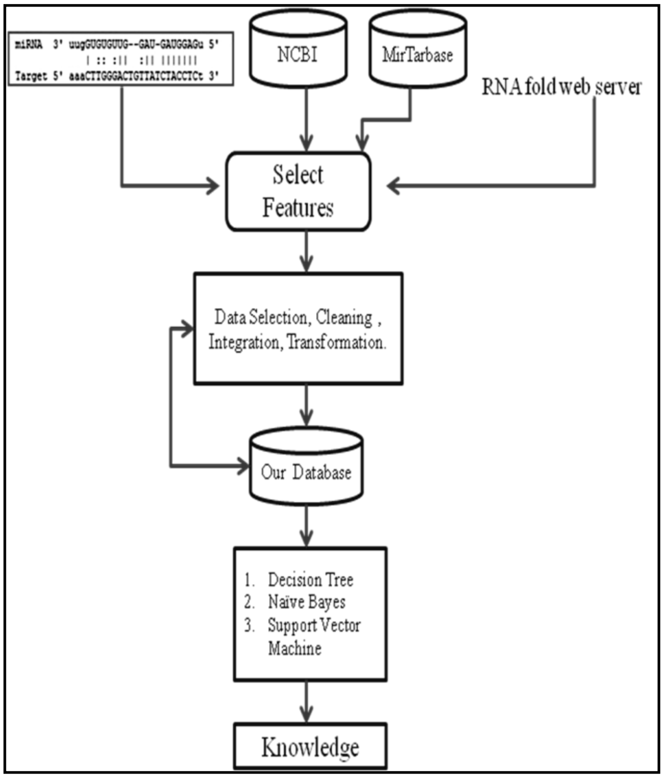
miRNA 3' uugGUGUGUUG--GAU-GAUGGAGu 5'
| :: :|| :|| ||||||||
Target 5' aaaCTTGGGACTGTTATCTACCTCt 3'
NCBI
MirTarbase
RNA fold web server
Select Features
Data Selection, Cleaning , Integration, Transformation.
Our Database
1. Decision Tree
2. Naïve Bayes
3. Support Vector Machine
Knowledge

Fig. 1. The research methodology.

B. Data Set

The result of this study depends on the quality of the datasets. Therefore, the miRNA features are collected from miRTarBase database that contain more than 3500 MTIs (miRNA–target interactions). The database content is updated manually by surveying pertinent literature using filter research articles related to functional studies of miRNAs. In general, reporter assays, western blot, or microarray experiments with over expression or knockdown of miRNAs validate experimentally the gathered MITLs.

miRTarBase currently contains 4,270 experimentally verified MTIs; it contains between 669 miRNAs and 2,533 target genes amongst 14 species. The miRTarBase provides the last updated collection compared with other similar miRNA databases. In addition, it contains the largest amount of validated MTIs [25].

miRTarBase database is now available on www.miRTarBase.mbc.nctu.edu.tw

The authors of [26] updated the research article content by continuously surveying research articles. The miRTarBase database depends on miRand algorithm of target prediction and gives three positions on the 3’UTR of mRNA for the same miRNA.

In addition, the target gene information that is collected from Entrez Gene is the database that contains gene-specific at the National Center for Biotechnology Information (NCBI), it is available on www.ncbi.nlm.nih.gov [27].

The result of curation and automated integration of data from NCBI's Reference Sequence project (RefSeq) is represented by the content of Entrez Gene that collaborates model organism databases, and that takes from many other databases available from NCBI. Records that are recommended are unique, stable, and tracked integers as identifiers. The content (i.e. nomenclature, map location, gene products and their attributes, markers, phenotypes, and links to citations, sequences, variation details, maps, expression, homologs, protein domains and external databases) becomes available by updating it as new information [28].

C. Data Selection

Table 1 contains all Attributes that are included in the final dataset. Justification for the censored attributes is shown in Table 2.

TABLE I. DATASET ATTRIBUTES

NO.	Attribute	Description
1	miRTarBase ID	The ID of miRNA–target interactions in miRTarBase
2	miRNA Name	microRNA Name
3	miRNA Position	first and end sites binding
4	miRNA Length	Length of microRNA Sequence
5	Target Gene Name	The RNA Name where the miRNA bind
6	mRNA Chromosome Number	Chromosome Number that contain the Target Gene
7	mRNA Chromosome location	The Location Of Target Gene on Chromosome
8	3'UTR Length	The sequence length of three prime untranslated region.
9	miRNA target sites Position	The start and end binding position of miRNA on target gene
10	Position	miRNA binding on target gene is F or M or E
11	Score	The Score of miRNA : mRNA target gene binding
12	miRNA Sequence	A set of characters that represent a miRNA Sequence
24	Number of Mismatch	the total number of (G:U) or(G:T) binding in miRNA: mRNA duplexes
25	Number of Bulges	the total number of not binding in miRNA: mRNA duplexes
26	Number of Loop	the total number of not binding and mead loop shape in miRNA: mRNA duplexes
27	Number of Match	the total number of binding in miRNA: mRNA duplexes

TABLE II. JUSTIFICATION FOR ATTRIBUTES CENSORED FROM THE FINAL DATASET

Attributes	Justification
G% of (miRNA), C% of (miRNA), U% of (miRNA), A% of (miRNA), # G of (miRNA), #C of (miRNA), # U of (miRNA), # A of (miRNA), C+G% of (miRNA)	Low number of miRNA included in the study.
miRNA Length	All miRNA in this study were 22 nt long.
MW of (miRNA), MW of (two stranded miRNA)	Low number of miRNA included in the study.
MFE	After the experiment had no effect on the results of classification

D. Data Cleaning

In this phase, we encountered three records out of 306 records included in this study that do not have information about target gene prediction. The action was to delete them from the dataset.

E. Data Integration

Many features were collected and integrated from miRTarBase and NCBI databases, and then they are considered as attributes for this research dataset. And the RNAfold web server is used [29] to compute (Minimal Free Energy) MFE feature of target gene, as shown in Table 1.

F. Data Transformation

All features that have" indirectly" value to compute another features such as the Position attribute are selected to be added. Position attribute contains three values: First (F), Middle (M) and End (E). Thus, the position represents the relative proximity of the binding site to the end of the RNA target regardless of the 3'UTR length. Dividing the 3'UTR into three regions was important to maintain a meaningful perspective of the analysis. These values (F, M and E) are calculated by dividing the length of 3' UTR of target gene on number three. Then look up where the miRNA target sites position to determine F, M or E. for example, length of 3'UTR for BCL2 target gene is 5282 and the miRNA target sites Position is 3377-3399, so the position value will be M.

IV. DATA MINING

A. The Final Dataset

After the transformation process, each kind of miRNA was separated with these attributes to prepare the final datasets ,those attributes are: miRNA Length, Target gene Name, miRNA target sites Position, Score, 3'UTR Length , number of mismatch, number of bulges, number of loop and number of match. These attributes were selected from miRTarBase database and the last four attributes were computed manually. In addition, another two attributes were taken from NCBI database: mRNA target gene Chromosome Number and mRNA target gene Chromosome Location. A sample of the final dataset is shown in table 3.

TABLE III. SAMPLE FROM LET-7A MIRNA DATASET AFTER TRANSFORMATION STEP

3'UTR Length	mRNA Chromosome Location	Position	Score	# of Mismatch	# of Blugs	# of Loop	# of Match
492	q	F	139	3	0	3	11
5282	q	M	144	3	3	1	13
1936	q	E	132	2	1	0	8
2510	q	M	164	3	2	1	14
3893	q	M	142	3	2	1	12
3640	P	F	148	3	5	0	11
4688	P	F	134	6	1	0	10
2456	q	M	146	4	2	2	13
471	P	M	154	1	3	1	13
1399	q	M	120	3	3	2	11
2376	q	M	156	2	2	2	13
3228	q	E	137	1	2	2	13

B. Classification Techniques

This study uses the following classification technique:

- **Support Vector Machine** (**SVM**): supervised learning machine tool that is used to classify a sample of data set into two predefined classes, based on statistical analysis [30].
- **Naive Bayes Classifier**: a simple supervised learning machine tool that employs Bayes' theorem with independence assumptions among features [31].
- **Decision Tree Learning**: supervised machine learning tool that is used as a predictive model to represent all effective (i.e. higher weight) decisions. Tree Leaves represent the possible classes while the edges represent conjunctions of features [32].

Step 1: Classification Using Target Gene Chromosome Location as a Class Label.

The data mining classification algorithms are applied for each kind of miRNA where the position attribute is the class label. However, the results are reported accuracies below 50%. Therefore, we made the mRNA target gene Chromosome Location attribute to be the class label, which started to provide improvement in the accuracy. Three algorithms of classification from the weka 3.6 software are used (decision tree –J48, naïve Bayes and support vector machine -SMO) where each algorithm is applied twice on each file with different ways of data training split: Cross-validation and percentage split, in which they are the default settings.

Step 2: Classification after Addition of Chromosome Categories to the Dataset

After that, a new feature is selected to add to the features set. The mRNA target genes Chromosome Number are grouped to four categories. Therefore, four files were made for each miRNA representing one of the categories to be as a class label with the above mentioned attributes.

The four categories of chromosomes classification are the following:

1) According to gene number: where the gene number is either lower or greater than 1000.

2) According to gene size: large, medium, or short.

3) According to gene satellites: whether the chromosome contains tandem repetitive DNA satellites sequence or not.

4) According to the mix between gene size, satellites and centromeric type: "Class" .

V. EXPERIMENT AND RESULTS

In this study, three algorithms in WEKA software are used (*e.g. decision tree –J48, naïve Bayes and support vector machine -SMO*) with default setting of data training split way, Cross-validation 10 folds or percentage split 66%.

In the first step, the aforementioned three algorithms were performed on all miRNA family datasets using the position attribute as a class label. Unfortunately, the accuracy of the results is lower than 50%. However, after changing the class label to the mRNA chromosome location attribute, the result has improved. Using dataset shown in table 3, we applied classification methods using miRNA position attribute with chromosome location attribute as a class label.

The highest accuracy that is reported using decision tree algorithm in Step 1 was for miR-21 (Acc=73.68%) and let-7a (Acc= 69.23%) in 66% test mode. Whereas, in the 10 fold test mode the highest accuracy was reported in miR-21 (Acc=65.45%) and miR-122 (Acc=60.47%) as shown in figure 2A. A clear difference between 66% and 10 fold test modes was only seen in let-7a by a shift of 20%. The highest accuracy that is reported using naïve Bayes in Step 1 was seen in miR-122 (Acc=60%) and miR-21 (Acc=57.89%) in 66% test mode. Whereas, in the 10 fold test mode the highest accuracy was seen in miR-21 (Acc=67.27%) and let-7 family (Acc=57.21%) as shown in figure 2B. Using the support vector machine algorithm almost, all miRNA families provide equal accuracies that are over 50% except for let-7 family in 66% testing mode. Whereas, in 10 fold test mode all accuracy values were equal and over 50% except the let-7a as shown in Figure 2C.

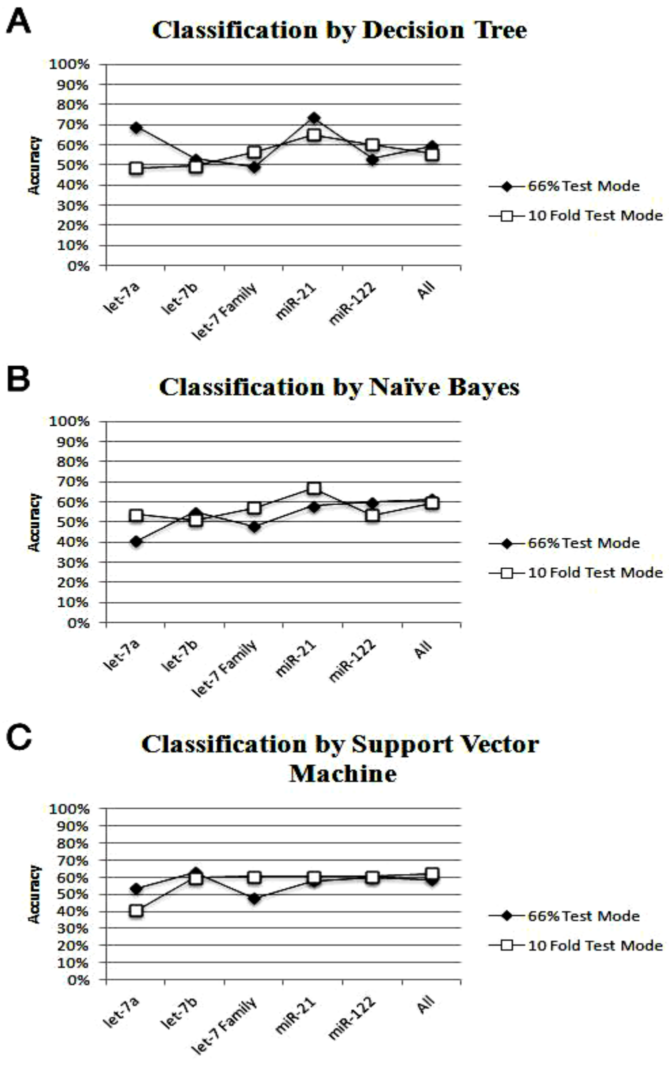

Fig. 2. The accuracy of miRNA-mRNA predictions in step 1 (before the addition of chromosome categorization) according to (A) Classification using decision tree. (B) Classification using naïve base. (C) Classification using support vector machine.

The highest accuracy that was reported using support vector machine was seen in let-7a (Acc=76.92%) and miR-21 (Acc=73.68%) in 66% test mode. Whereas, in 10 fold test mode the highest accuracy was seen in let-7a (Acc=79.49%) and miR-21 (Acc=69.09%) as shown in Figure 5C.

In general, accuracy was higher in step 2 when it is compared to step 1 using any of the three algorithms as shown in Figure 3.

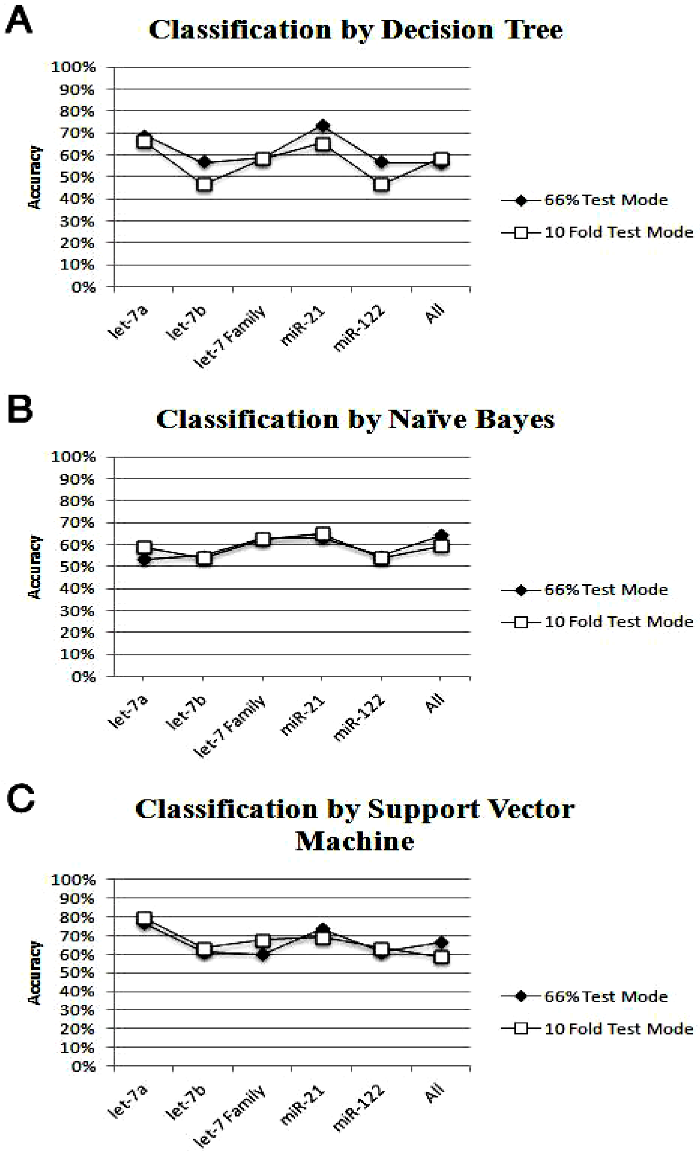

Fig. 3. The accuracy of miRNA-mRNA predictions in step 2 (after the addition of chromosome categorize) according to (A) Classification using decision tree. (B) Classification using naïve base. (C) Classification using support vector machine.

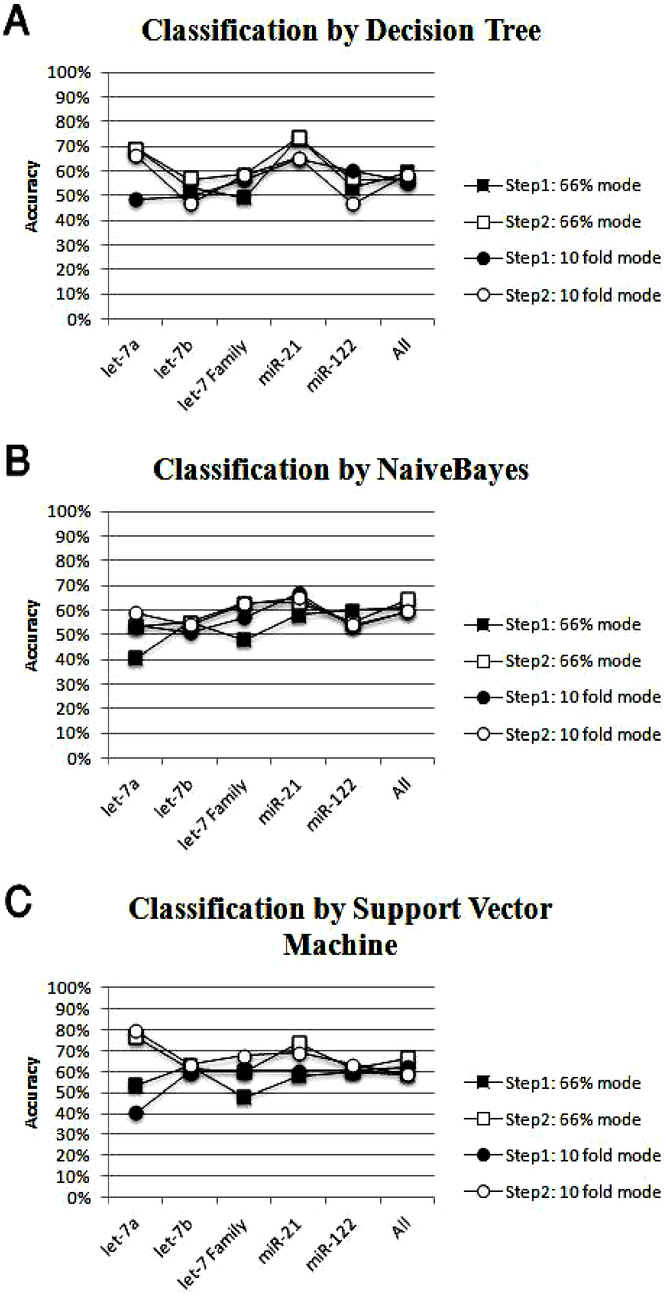

Fig. 4. The accuracy of miRNA-mRNA predictions showing differences between step 1 (before the addition of chromosome categorize) and step 2 (after the addition of chromosome categorize) according to (A) Classification using decision tree. (B) Classification using naïve base. (C) Classification using support vector machine.

In order to compare the different ways of chromosome categorization and analyze their accuracy in miRNA-mRNA prediction, four categorization methods were applied. Using the decision tree algorithm in 66% test mode, the highest accuracy in the four categorizations was reported in miR-21 (Acc= 73.68% in all categorization methods) as shown in Figure 5A. In addition, let-7a provides high accuracy in all methods (Acc=76.92%, 69.23%, 69.23%, in satellite categorization, size categorization, class categorization, respectively) except for the gene number categorization (Acc= 46.15%).

Using the naïve Bayes algorithm in 66% test mode, the accuracy in the four categorizations was reported between 46.15% to 61.54% and did not provide clear differences between miRNA families or chromosome categorizations as shown in Figure5B.Using the support vector machine algorithm in 66% test mode, the class categorization was clearly showing the highest accuracy as shown in Figure 5C.

Using the decision tree algorithm in 10 fold test mode, the highest accuracy in the four categorizations was reported in miR-21 using gene number categorization (Acc=76.36%) followed by let-7a using class categorization (Acc=66.67 %) as shown in Figure 6A.

Using the naïve Bayes algorithm in 10 fold test mode, the highest accuracy in the four categorizations was reported in miR-21 between Acc=61.82% to Acc=69.09% and all miRNA between Acc=58.82% to Acc=64.71% as shown in Figure 6B.

Using support vector machine in 10 fold test mode, the class categorization was clearly showing the highest accuracy (Figure 6C). Interestingly, let-7a showed the highest accuracy in class categorization (Acc=79.49%) and rather the lowest accuracy for the rest of categorizations (Acc=38.46% to Acc=51.28).

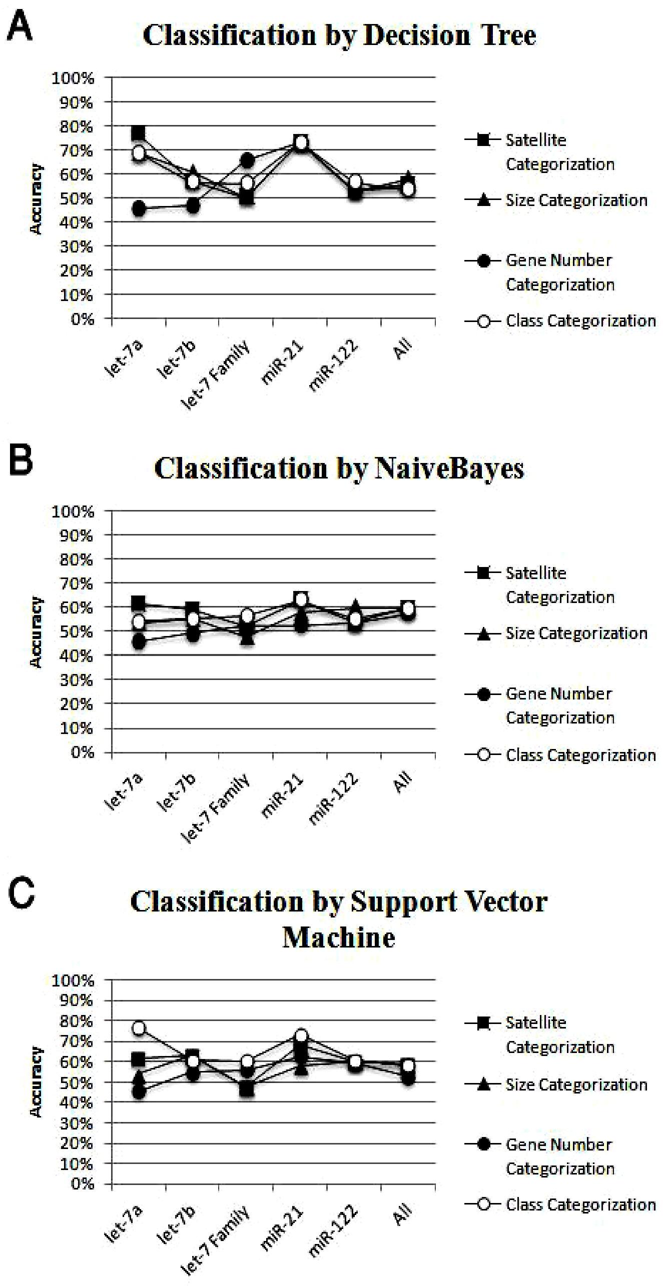

Fig. 5. The accuracy of miRNA-mRNA predictions in 66% test mode showing differences between four categorizations in step 2 according to (A) Classification using decision tree. (B) Classification using naïve base. (C) Classification using support vector machine.

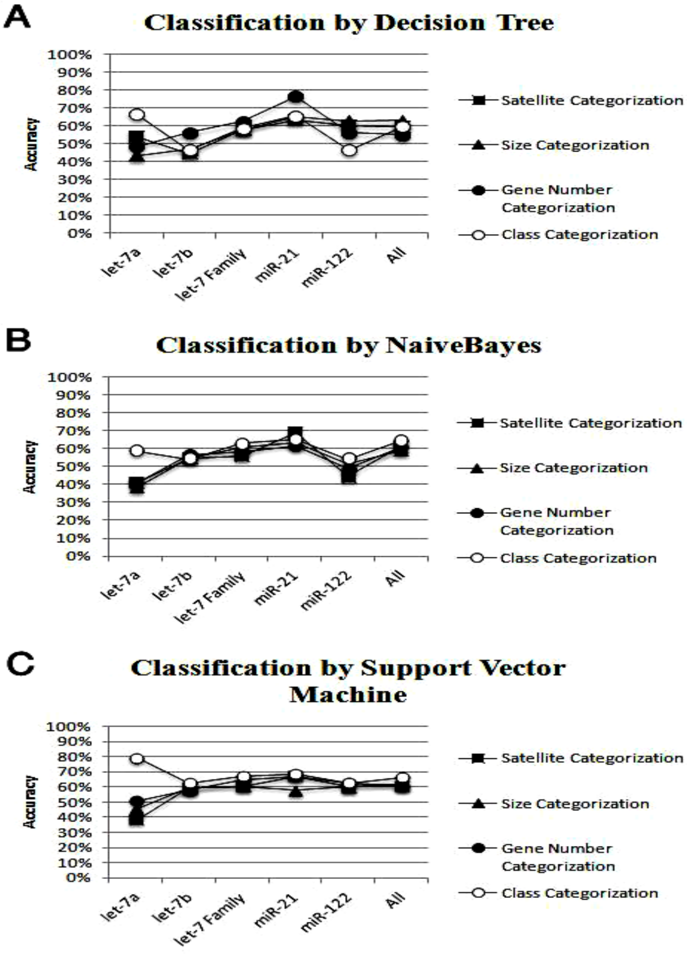

Fig. 6. The accuracy of miRNA-mRNA predictions in 10 fold test mode showing differences between four categorizations in step 2 according to (A) Classification using decision tree. (B) Classification using naïve base. (C) Classification using support vector machine.

Using chromosome class as a categorization method, the highest accuracy reported using decision tree was seen in miR-21 (Acc=73.68%) and let-7a (Acc=69.23 %) in 66% test mode. Whereas, in 10 fold test mode the highest accuracy was seen in let-7a (Acc=66.67%) and miR-21 (Acc=65.45%) as shown in Figure 2A. The highest accuracy that is reported using naïve Bayes was seen in All miRNA (Acc=64.71%) and miR-21 (Acc=63.16%) in 66% test mode. Whereas, in 10 fold test mode, the highest accuracy was seen in miR-21 (Acc=65.45%) and let-7 family (Acc=62.98%) as shown in Figure 2B.

A. Results Discussion

In this study, we focused on finding a correlation between the miRNA target sites of specific types of miRNA, mainly; let-7a, let-7b, let-7 family, miR-21 and miR-122 and the chromosomal location, the nucleotide sequence, binding and thermodynamic features of miRNA and mRNA. The dataset was collected from miRNA target prediction database and then three techniques of data mining (*i.e. decision tree –J48, naïve Bayes and support vector machine -SMO*) were used to investigate the correlations by weka 3.6 software with default setting of data training split way, Cross-validation 10 folds or percentage split 66%.

Data mining procedure was performed in two steps. The first step was done without adding the chromosome categorization and the second step was done with adding of any of four different chromosome categorization methods.

In general, the accuracy was higher in step 2 when it is compared with step 1 using any of the three algorithms as provided in Figure 6.

When using the decision tree in Step 1, the highest accuracy reported was for miR-21 and let-7a in 66% test mode and for miR-21 and miR-122 in the 10 fold test mode as appears in Figure 4A. In step 2, the highest accuracy reported was seen in miR-21 and let-7a in both testing modes as appears in Figure 5A. In terms of chromosome categorization, the miR-21 reported the highest accuracy in all categorization methods using the 66% test mode as appears in Figure 5A and the highest accuracy using gene number categorization in the 10 fold test mode as appears in Figure 6A. It is clear that miRNA-21 had the best accuracy, while the let-7a provided a clear decrease in the accuracy in the 10 fold test mode by 20% which might be attributed to the nature of the records that are included in the training and testing datasets. When using naïve Bayes in Step 1, the highest accuracy reported was for miR-21 and miR-122 in 66% test mode and for miR-21 and let-7a in the 10 fold test mode as appears in Figure 4B. These results were similar to the results shown in the decision tree. In step 2, the highest accuracy was seen in “All miRNA” and miR-21 in 66% test mode and in miR-21 and let-7 family in the 10 fold mode as appears in Figure 5B. Thus far, miR-21 was predominantly showing the highest accuracy in both decision tree and naïve Bayes. Additionally, let-7a provided a high accuracy level in all categorization methods (Acc=76.92%, 69.23%, 69.23%, in satellite categorization, size categorization, class categorization, respectively) except for gene number categorization (Acc=46.15 %). It is possible that the gene number categorization has this drop in accuracy due to the imbalanced number of records in let-7a (9 below 1000 genes vs. 30 over 1000 genes). In general, the 66% test mode for all categorizations reported accuracy in a low range (46.15% to 61.54%) with no clear differences between miRNA families or chromosome categorizations as appears in Figure 5B. Whereas, the 10 fold test mode for all categorizations showed the highest accuracy in the miR-21 and “All miRNA” as appears in Figure 6B. Naïve Bayes is a probabilistic statistical method that can be easily affected by the frequency of the attributes and the number of records, which again is why it was not a surprise when “All miRNA” dataset were providing a highest accuracy in many testing cases.

When using support vector machine in step 1, almost all miRNA families showed similar accuracies over 50% except for let-7 family in 66% testing mode and let-7a in 10 fold testing mode as appears in Figure 4C. In step 2, the highest accuracy reported was seen in let-7a and miR-21 in 66% test mode and let-7a and miR-21 in 10 fold test mode as appears in Figure 5C. The class categorization was clearly providing the highest accuracy in both testing modes as appears in Figure 5C and Figure 6C. Interestingly, in the 66% test mode, let-7a showed the highest accuracy in class categorization (Acc=79.49%) and rather the lowest accuracy for the rest of categorizations between (Acc=38.46% to Acc=51.28). Support vector machine is a non-probabilistic machine learning method that employs the addition of a hyperplane or more (i.e. extra dimension or dimensions). In principle, when the attributes are categorized in a larger number of groups, the algorithm gains more freedom to construct further hyperplanes (i.e. more dimensions). This can be used to explain the let-7a case where class categorization is composed of 7 groups (i.e. allowing for more dimensions). Whereas, the rest of the categorizations are composed of 2-3 groups.

Eventually, this study provides an outline of the major factors involved in miRNA-mRNA target prediction. Out of 26 features included in this study, only 9 features were retained. The rest of the features were eliminated mostly due to the low number of miRNAs included in this study except for one attribute (i.e. MFE) which had no effect ,whatsoever, on the results. MFE and the score were the only thermodynamic parameters in the study and it might be better to use more complex thermodynamic parameters in the future. The factors that were involved in prediction including match, mismatch, bulge, loop, and score represent the binding characteristics, while the position, 3’UTR length, and chromosomal location and chromosomal categorizations represent the characteristics of the target mRNA. Several attributes such as the match, mismatch, bulge, and 3’UTR length can provide empirical guidance for study of specific miRNAs using decision trees because they are classified according to an optimized cutoff value (i.e. a threshold) which cannot be inferred experimentally. In addition, some of the attributes such as the chromosomal location and chromosomal categorization have never been studied before as factors of prediction and yet they have been shown here to play a major role in the prediction. The chromosome location was a class label in this study, while the chromosomal categorizations provided the increased accuracy in prediction in all three algorithms, suggesting that they might have a major biological influence by controlling the gene expression of different cellular pathways.

VI. Conclusion and Future Work

miRNA research has been developed progressively in the past few years. Prediction of miRNA-mRNA target was attempted both computationally and experimentally. In this study, data mining techniques were used to classify a number of characteristics involved in miRNA binding and the mRNA targets themselves. Five families of miRNAs that are involved in cancer pathways have been analyzed in this study. The results can be summarized as follows:

The use of decision tree in miRNA-mRNA target prediction shows that each miRNA family behaves in a unique way when it comes to binding features with or without chromosomal categorization:

- The highest accuracy reported without chromosomal categorization was for miR-21 and let-7a in 66% test mode and for miR-21 and miR-122 in the 10 fold test mode and with chromosomal categorization the highest accuracy reported was seen in miR-21 and let-7a in both testing modes.

- The decision tree of let-7a showed the greatest weight to the mismatch followed by position and score without chromosomal categorization. While in step 2 the class categorization became the root for the tree followed by the matches and bulges.
- The decision tree of miR-21 was the same with and without chromosomal categorization. The root was the match attribute followed by 3'UTR length.
- The decision tree of miR-122 shows a clear complication and branching when the class categorization was added. The tree develops from one weight attribute in step 1 which was the 3'UTR length into a more complicated branching in step 2 including many attributes.
- Binding features such as the match, mismatch, and bulge as well as the length of the 3'UTR was shown to play major role in the classification of targets. In addition, the chromosomal source of the target that is represented here by the class categorization contributed in the accuracy of the test.

The use of naïve Bayes without chromosomal categorizations showed the highest accuracy in miR-21 and miR-122 families in 66% test mode and for miR-21 and let-7a in the 10 fold test mode. Whereas, when the chromosomal class categorization was used, the highest accuracy was seen in "All miRNA" and miR-21 in 66% test mode and in miR-21 and let-7 family in the 10 fold mode.

When using support vector machine without chromosomal categorization almost all miRNA families showed similar accuracies(over 50%) except for let-7 family in 66% testing mode and let-7a in 10 fold testing mode. When chromosomal categorization was used, the highest accuracy reported was seen in let-7a and miR-21 in 66% test mode and let-7a and miR-21 in 10 fold test mode.

Out of 26 features included in this study, only 9 features were retained. The rest of the features were eliminated either due to the low number of miRNAs included in this study or because they did not have any effect on the experimental results. The factors that were involved in prediction including match, mismatch, bulge, loop, and score represent the binding characteristics, while the position, 3'UTR length, and chromosomal location and chromosomal categorizations represent the characteristics of the target mRNA.

In the future, several attributes such as the match, mismatch, bulge, and 3'UTR length can provide a threshold-based empirical guidance for study of specific miRNAs to scan the whole human genome for novel targets.

In addition, naïve Bayes and support vector machine can be used to test the new attributes, especially the ones involved in the source of the target mRNA (i.e. chromosomal based attributes).

New findings in the field of miRNA have the potential to revolutionize the study of many diseases. Many of the known miRNA are under focus now for targeted medicine and research is now ongoing in the field of using these miRNA as drugs for treatment of different types of cancer and diagnosis.

References

[1] James Watson, Tania Baker, Stephen Bell, Alexander Gann, Michael Levine, Richard Losick. (2003) Molecular Biology of the Gene, Fifth Edition, Pearson (Benjamin Cummings) Publishing.

[2] Mohammed Abba, Heike Allgayer. MicroRNAs as regulatory molecules in cancer: a focus on models defining miRNA functions. Drug Discovery Today: Disease Models 2009; 6(1): 13-19

[3] Jürgen Wittmann, Hans-Martin Jäck. Serum microRNAs as powerful cancer biomarkers 2010; 1806: 200–207.

[4] Erik A.C. Wiemer. The role of microRNAs in cancer: No small matter. Euro pean journal of cancer 2007; 43: 1529 –1544.

[5] Cathy A. Andorfer, Brian M. Necela, E. Aubrey Thompson and Edith A. Perez. MicroRNA signatures: clinical biomarkers for the diagnosis and treatment of breast cancer. Trends in Molecular Medicine 2011; 17(6): 313-319.

[6] Haverich, Carina Gross, Stefan Engelhardt, Georg Ertl, Johann Bauersachs van Laake, Pieter A. Doevendans,el at. MicroRNAs in the Human Heart : A Clue to Fetal Gene Reprogramming in Heart Failure. American Heart Association 2007; 116: 258-267.

[7] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. From Data Mining to Knowledge Discovery in Databases. AI MAGAZINE 1996; 17(3): 37-54.

[8] Pang-Ning Tan,Michael Steinbach and Vipin Kumar. Introduction to Data Mining. International Edition. Boston 2006; ISBN-10: 0321420527

[9] Hasan Og˘ul, Sinan U. Umu, Y. Yener Tuncel and Mahinur S. Akkaya. A probabilistic approach to microRNA-target binding. Biochemical and Biophysical Research Communications 2011; 413: 111-115.

[10] Bing Liu, Jiuyong Li and Anna Tsykin. Discovery of functional miRNA–mRNA regulatory modules with computational methods. Journal of Biomedical Informatics 2009; 42: 685–691.

[11] Wan Hsieh, Hsiuying Wang. Human microRNA target identification by RRSM. Journal of Theoretical Biology 2011; 286: 79–84.

[12] National Chiao-Tung University [Internet]. Taiwan: National Chiao-Tung University; Available from: http://www.stat.nctu.edu.tw/_hwang/website_wang%20new.htm

[13] William Ritchie, Megha Rajasekhar, Stephane Flamant, John Rasko. Conserved Expression Patterns Predict microRNA Targets. PLoS Computational Biology 2009; 5(9).

[14] Xiaofeng Song , LeiCheng , TaoZhou , XuejiangGuo , XiaobaiZhang , el at. Predicting miRNA-mediated gene silencing mode based on miRNA-target duplex features. Computers in Biology and Medicine 2011; 42(1): 1-7.

[15] Scott T. Younger , Alexander Pertsemlidis , David R. Corey. Predicting potential miRNA target sites within gene promoters. Bioorganic & Medicinal Chemistry Letters 2009; 19: 3791–3794.

[16] Alain Sewer, Nicodème Paul, Pablo Landgraf, Alexei Aravin, el at. Identification of clustered microRNAs using an ab initio prediction method. BMC Bioinformatics 2005, 6:267.

[17] Xingqi Yan, Tengfei Chao, Kang Tu, Yu Zhang, Lu Xie, Yanhua Gong, Jiangang Yuana, Boqin Qiang and Xiaozhong Peng. Improving the prediction of human microRNA target genes by using ensemble algorithm. FEBS Letters 2007; 581: 1587–1593.

[18] Memorial Sloan-Kettering Cancer Center. Memorial and associates; 2005 Available from: http://www.microrna.org/microrna/getMirnaForm.do

[19] Malik Yousef, Segun Jung, Andrew V. Kossenkov, Louise C. Showe and Michael K. Showe. Naıve Bayes for microRNA target predictions—machine learning for microRNA targets. Bioinformatics 2007; 23(22): 2987–2992.

[20] University of Pennsylvania education [Internet]. Philadelphia: University of Pennsylvania and associates; 2007 [updated 2012 November 1]. Available from: http://wotan.wistar.upenn.edu/NBmiRTar/

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

[22] Chenghai Xue, Fei Li, Tao He1, Guo-Ping Liu, Yanda Liand Xuegong Zhang. Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine. BMC Bioinformatics 2005; 6

[23] Yunfei Pei, Xi Wang, Xuegong Zhang. Predicting the fate of microRNA target genes based on sequence features. Journal of Theoretical Biology 2009; 261: 17-22.

[24] M. Sualp, T. Can. Using network context as a filter for miRNA target prediction. BioSystems 2011; 105: 201-209.

[25] Sheng-Da Hsu, Feng-Mao Lin, Wei-Yun Wu, Chao Liang, Wei-Chih Huang, Wen-Ling Chan, et al. miRTarBase: a database curates experimentally validated microRNA–target interactions. Nucleic Acids Research 2011;39: 163-169

[26] National Chiao Tung University [Internet].Taiwan: National Chiao Tung University and Association; [updated 2011 October 15; cited 2011 december 25] Available from: http://miRTarBase.mbc.nctu.edu.tw/

[27] National Center for Biotechnology Information [Internet]. U.S: National Library of Medicine and Association; Available from: www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene

[28] Donna Maglott, Jim Ostell, Kim D. Pruitt and Tatiana Tatusova. Entrez Gene: gene-centered information at NCBI. Nucleic Acids Research 2005; 33: 54-58.

[29] University of Vienna. University of Vienna and Association; Available from: http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi

[30] Y.H. Qiao, J.L. Liu, C.G. Zhang, X.H. Xu and Y.J. Zeng. SVM classification of human intergenic and gene sequences. Mathematical Biosciences 2005; 195: 168-178

[31] Francesca Demichelis, Paolo Magni, Paolo Piergiorgi, Mark A Rubin and Riccardo Bellazzi. A hierarchical Naïve Bayes Model for handling sample heterogeneity in classification problems: an application to tissue microarrays. BMC Bioinformatics 2006; 7: pages numbers

[32] Shuyan Chen, Wei Wang. Decision tree learning for freeway automatic incident detection. Expert Systems with Applications 2009; 36: 4101–4105.

[33] Behzad Rabiee-Ghahfarrokhi, Fariba Rafiei, Ali Akbar Niknafs, Behzad Zamani, "Prediction of microRNA target genes using an efficient genetic algorithm-based decision tree", FEBS Open Bio 2015; 5: 877–884