Feature based Analysis and Classification of Plant MicroRNAs
Alma Dabral1, Saad Murtaza Khan2, A. K. Mishra3 and H. Chandrasekharan4
1,2,3,4Unit of Simulation and Informatics, Indian Agricultural Research Institute, New Delhi 110012
1almadabral@gmail.com, 3akmisha@iari.res.in, 2saadmurtazakhan@gmail.com and 4head_usi@iari.res.in

ABSTRACT
Studies on microRNA (miRNA) have progressed tremendously in recent past, but further computational analysis is required to know the complete potential of these non coding RNAs. Due to its short length (~20 nucleotides), it is difficult to use the conventional genetic techniques for miRNA prediction and analysis. This has led to the computational analysis of miRNAs. These are non-coding small RNAs which are responsible for gene regulation at the post translational level by binding to the miRNAs and thereby stopping the translation activities. In this paper we have studied about 1000 miRNA and precursor miRNA sequences from monocots (Oryza sativa, Zea mays, Sorghum bicolor) and Brassicaceae (Arabidopsis thaliana and Brassica napus). Our study in this paper is on the miRNA classification using decision trees. Some dominating attributes were derived for the analysis of the miRNA sequences. We have used WEKA (a data mining tool), which helps us to study the large data for its classification. The decision trees based classifications are best suited for the analytical study of miRNA and the derived dominating attributes are biologically significant.

KEYWORDS
miRBase, miRNA, precursor, RNAFold, WEKA, J48, Decision Trees

1. INTRODUCTION
MicroRNAs (miRNA) are small non coding RNA molecules which are 18-25bp long sequences. They are responsible for gene regulation in plants as well as in animals. Present in the nucleus, miRNA genes are transcribed into primary transcript or pri-miRNA with the help of RNA polymerase-2. The dsRNA specific ribonuclease Drosha digests the long primary miRNA transcript in the nucleus and releases hairpin precursor miRNA (pre-miRNA). Exportin-5(Exp5) and RAN-GTP transports the pre-miRNA into the cytoplasm where an enzyme called Dicer, processes the pre-miRNA into mature miRNA [1, 2]. Dicer (endonuclease) is a member of Rnase-3 superfamily and cleaves the pre-miRNA approximately 19bp from the Drosha cut site. Only one of the two strands is the miRNA. The double stranded RNA produced by Dicer separate and associate with RISC (RNA-induced silencing complex)[3, 4], based on the stability of the 5’end. Plants lack Drosha therefore Dicer performs the processing [5..7].

Mature miRNA are partially complimentary to messenger RNA and they play an important role in gene regulation through mRNA cleavage or translational repressal by associating with the RISC. Prediction of miRNA helps us to understand its structure and therefore its function and role in organism [8, 9]. In plants, they regulate the development of leaves and flower. Thus intense study is required to find out the regulation of most fundamental biological processes in the organism.

1.1 Need for computational prediction
The short length of miRNA makes it difficult to analyse it with the help of conventional genetic techniques. Some miRNAs have low expression levels and some are expressed in specific conditions only, due to this reason their cloning is difficult. Also Deep-sequencing techniques require intense computational analysis to differentiate the miRNAs from other non-coding miRNAs [10, 11]. Therefore we look up to the computational approaches to predict miRNA sequences and do their analysis.
Due to the short length of miRNA sequences, tools like BLAST give a large number of irrelevant hits. Hence only nearly perfect matches are to be found. Also the pre-miRNA sequences are less conserved which makes it difficult to use the conventional sequence alignment methods to find the homologous. Unlike the sequences, the secondary structures are more conserved which is helpful in predicting new miRNAs. Therefore more sensitive methods which consider both sequence and structure conservation are needed [12..15].

2. LITERATURE SURVEY
To carry out the computational prediction and their analysis there are some known tools, which are based on the following algorithms [16..18].
- Filter based- This approach uses different features and conservation criteria to restrict the precursor candidates [19, 20].
- Machine learning- It uses the concept of learning through previously known miRNAs.
- Mixed approach- In this a combination of computational tools and high-throughput experimental procedures are used.
- Target centered approach- From conservation analysis a putative set of miRNA targets are developed which helps to find out new miRNAs.
- Homology based- Identifies the miRNAs similar to previously known pre-miRNAs.
- Rule based- It is based on some rules by studying the features of the sequences.
We downloaded the set of miRNAs and precursor miRNAs from miRBase (version 15) [21]. It has 1010 known mature miRNA sequences from 5 species; Oryza sativa (447), Arabidopsis thaliana (199), Zea mays (170), Sorghum bicolor (148) and Brassica napus (46). Oryza sativa and Arabidopsis thaliana are more in number as their genome sequence information is available.

### Reference miRNAs
The free and open sources for miRNA data are limited, but miRBse is a database which is open source and can be easily accessed for our studies. We downloaded the set of miRNAs and precursor miRNAs from miRBase (version 15) [21]. It has 1010 known mature miRNA sequences from 5 species; Oryza sativa (447), Arabidopsis thaliana (199), Zea mays (170), Sorghum bicolor (148) and Brassica napus (46). Oryza sativa and Arabidopsis thaliana are more in number as their genome sequence information is available.

### Preparation of dataset
Using PERL scripts we automated the retrieval of 1010 sequences from miRBase repository. These sequences were put into the RNAfold, software developed by M. Zuker and P. Stiegler, is a window based utility of Vienna RNA secondary structure server available at http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi. RNAfold calculates the minimum free energy structure for each sequence [22]. Coding in PERL was done to calculate the values of the set attributes from their secondary structures.

### Computational analysis
WEKA (Waikato Environment for Knowledge Analysis) is a JAVA based software developed at the University of Waikato, New Zealand. WEKA version 3.6.2 was used to do our research. It is a data mining tool written in java language which is a collection of machine learning algorithms. We are using the J48 classifier to classify our data as it is the easiest and simplest way to interpret the results [23].

### Data Curation
The 1010 miRNA sequences were downloaded from miRBase (15 release) with the help of Perl script. RNAfold was run on all the sequences and a secondary structure was generated along with the MFE of the sequences. The secondary structure is in the form of dot bracket format. Each bracket represents a base pairing and each dot a non paired base.

A mature miRNA sequence:-
>osa-miR395s MIMAT0000968
GUGAAPUGUUUUGGGGGAACUC

A precursor miRNA sequence:-
>osa-MIR395s MI0001037
GUAUCACGGAGAGGUUCUCUUCAAGACUCUCAGUUGGACACUUUCAGACUCUCUGUGGACACUCAGAUGUCC

#### Table 1: In the following table some widely used and popular tools are given: W-web based; D-downloadable

<table>
<thead>
<tr>
<th>Name of the tool</th>
<th>Type</th>
<th>Techniques</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mir Scan</td>
<td>W</td>
<td>Filter-based</td>
<td><a href="http://genes.mit.edu/mirscan/">http://genes.mit.edu/mirscan/</a></td>
</tr>
<tr>
<td>MiRFinder</td>
<td>D</td>
<td>Filter-based</td>
<td><a href="http://www.bioinformatics.org/mirfinder/">http://www.bioinformatics.org/mirfinder/</a></td>
</tr>
<tr>
<td>ProMIR</td>
<td>W</td>
<td>Machine learning</td>
<td><a href="http://cbit.snu.ac.kr/_ProMiR2/">http://cbit.snu.ac.kr/_ProMiR2/</a></td>
</tr>
<tr>
<td>TripletSVM</td>
<td>D</td>
<td>Machine learning</td>
<td><a href="http://bioinfo.au.tsinghua.edu.cn/mirnasvm/">http://bioinfo.au.tsinghua.edu.cn/mirnasvm/</a></td>
</tr>
<tr>
<td>MiPred</td>
<td>W</td>
<td>Machine learning</td>
<td><a href="http://www.bioinf.seu.edu.cn/mirRNA/">http://www.bioinf.seu.edu.cn/mirRNA/</a></td>
</tr>
<tr>
<td>Mireval</td>
<td>W</td>
<td>Mixed approaches</td>
<td><a href="http://tage.univ-mrs.fr/mireval/">http://tage.univ-mrs.fr/mireval/</a></td>
</tr>
<tr>
<td>findMiRNA</td>
<td>D</td>
<td>Target based</td>
<td><a href="http://sundarlab.ucdavis.edu/mirna/downloadd.html">http://sundarlab.ucdavis.edu/mirna/downloadd.html</a></td>
</tr>
<tr>
<td>MirAlign</td>
<td>W</td>
<td>Homology-based</td>
<td><a href="http://bioinfo.au.tsinghua.edu.cn/miralign/">http://bioinfo.au.tsinghua.edu.cn/miralign/</a></td>
</tr>
<tr>
<td>BayesiMiRN</td>
<td>W</td>
<td>Rule based</td>
<td><a href="http://wotan.wistar.upenn.edu/mirNA/">http://wotan.wistar.upenn.edu/mirNA/</a></td>
</tr>
</tbody>
</table>

### Identification of attributes and calculating the values
From the sequences, 9 and 14 attributes were derived for 1010 mature microRNA sequences and precursor sequence analysis respectively. Attributes for mature microRNA are ARM sequence on first or second arm of the hairpin structure, DFL distance of the mature sequence from loop sequence, BPN base pair per nucleotide, LNM length of the mature miRNA sequence, POP percentage of pairing, GCC Guanine and Cytosine nucleotide content in the sequence, MFE minimum free energy to fold the mature microRNA, DAS dominating nucleotide at start of the sequence and DAE dominating nucleotide at end of the sequence.

Attributes for precursor sequences are LEN length of the precursor sequence, NBP number of base pairs in the sequence, BLR base length ratio, NHP number of hairpins, HPL hairpin length, FRE free energy(minimum) to fold the sequence, FEN free energy(minimum) per nucleotide, AUC Adenine and Cytosine nucleotide content in the sequence, MTL maximum bulge in the sequence, SDI symmetric difference, MBL maximum length of the bulge, MBS maximum bulge symmetry, MTL maximum number of tails and NTL number of tails [24].

There are 9 attributes chosen for the mature microRNA:-
\[A1=\{ARM, DFL, BPN, LNM, POP, GCC, MFE, DAS, DAE\}\]

There are 14 attributes chosen for the precursor microRNAs:-
\[A2=\{LEN, NBP, BLR, NHP, HPL, FRE, FEN, AUC, MSK, SDI, MBL, MBS, MTL, NTL\}\]
Based on these attributes the values were calculated with the help of Perl scripts and datasets were prepared for different species.

**Attributes**

\[A = \{\text{LEN, NBP, BLR, NHP, HPL, FRE, FEN, AUC, MSK, SDI, MBL, MBS, MTL, NTL}\}\]

**Attribute values**

\[\{87, 33, 0.37, 1, 7, -37.20, 0.42, 56, 17, 2, 3, 6, 0, 0\}\]

To feed in the calculated data of the attributes of miRNAs, it was converted into ARFF format. Shuffle DNA was used to shuffle the precursor sequences of all the species in such a way that we generated sequences having hairpins. Using Perl script, randomly mature sequences were picked from the new randomised sequences and a negative dataset was created.

**4. RESULT AND DISCUSSION**

In this study, we find out the dominating attributes of the existing microRNAs of the plant species. In addition, decision trees were constructed of all the miRNA sequences by using a classifier. The graphical view represents some patterns found in the microRNA sequences. The dominating nucleotide at the beginning of the sequences is Uracil whereas Adenine has the lowest percentage. Near the end of the sequences Cytosine percentage is highest and Adenine percentage is lowest. The data shows that the mature microRNA sequences are mostly found on the first arm on the hairpin loop. Maximum number of mature miRNA has minimum distance from the hairpin thus restating the fact that the mature microRNA sequences are to be found near the hairpin loop. Maximum precursor sequence shows no tail in the secondary structures. The AU content is 60% in maximum number of precursor sequences and GC content is 62% in maximum number of mature miRNA sequences. The free energy to fold the precursor sequences was found mostly between -52Kcal/mol and -37Kcal/mol. The hairpin length was between 4 to 6 nucleotides long in maximum sequences. Presence of mostly single hairpin was found though there were cases of more than one hairpin loop in precursor sequences. There were sequences found having six hairpins which were considered under special occurrences.

We tested the data as a training set and generated the decision trees for the species **Oryza sativa** (447), **Arabidopsis thaliana** (199), **Zea mays** (170), **Sorghum bicolor** (148) and **Brassica napus** (46).

- **Decision tree construction**

The datasets mentioned in the previous section were fed into the Weka package for decision tree construction. A graphical representation of the dataset was displayed. This graphical view indicates various patterns in the dataset values. WEKA revealed that there are some attributes which are dominating than rest of the attributes. Weka provides us with attribute evaluators and search methods which help us find the dominating attributes. These attributes vary with different search methods and the evaluators. A tabular view of selected attributes is shown below.

<table>
<thead>
<tr>
<th>Attribute evaluator and search methods</th>
<th>Oryza sativa</th>
<th>Arabidopsis thaliana</th>
<th>Zea mays</th>
<th>Sorghum bicolor</th>
<th>Brassica napus</th>
</tr>
</thead>
<tbody>
<tr>
<td>BestFirst+CostSubsetEval</td>
<td>BLR, NHP,</td>
<td>NBP, BLR, NHP, FRE</td>
<td>BLR, NHP</td>
<td>BLR, FEN, NHP,</td>
<td>BLR, FEN, SBR, MBA</td>
</tr>
<tr>
<td></td>
<td>FRE, FEN</td>
<td></td>
<td>FRE</td>
<td>MSK</td>
<td></td>
</tr>
<tr>
<td>Ranker+ChiSquaredAttributeEval</td>
<td>BLR, NHP,</td>
<td>FEN, BLR, DFL, NHP</td>
<td>BLR, NHP</td>
<td>MBS, BLR, FEN,</td>
<td>MBS, FEN, BLR, SDI</td>
</tr>
<tr>
<td></td>
<td>FRE, FEN</td>
<td></td>
<td>FRE, MSK</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GreedyStepwise+ConsistencySubsetEval</td>
<td>BLR, NHP,</td>
<td>NBP, BLR, NHP, FRE</td>
<td>BLR, NHP</td>
<td>BLR, MBS, BLR,</td>
<td>BLR, MBS</td>
</tr>
<tr>
<td></td>
<td>FRE, FEN</td>
<td></td>
<td>FRE, MSK</td>
<td>MBA</td>
<td></td>
</tr>
<tr>
<td>Ranker+SVMainAttributeEval</td>
<td>BLR, NHP,</td>
<td>FEN, BLR, DFL, NHP</td>
<td>BLR, NHP</td>
<td>MBS, MBA, BLR,</td>
<td>MBS, FEN, BLR, MBA</td>
</tr>
<tr>
<td></td>
<td>DFL, FEN</td>
<td></td>
<td>DFL, FRE</td>
<td>MSK</td>
<td></td>
</tr>
<tr>
<td>Ranker+InfoGainAttributeEval</td>
<td>BLR, NHP,</td>
<td>FEN, BLR, DFL, NHP</td>
<td>BLR, NHP</td>
<td>MBS, BLR, FEN,</td>
<td>MBS, FEN, BLR, MBA</td>
</tr>
<tr>
<td></td>
<td>FRE, DFL</td>
<td></td>
<td>FRE, MSK</td>
<td>SDI</td>
<td></td>
</tr>
</tbody>
</table>

**Table 2**: Selected attributes (20)

By observing the graphs certain patters are recorded of the different species. The classifier J48 which is an implementation of C4.5 algorithm works on the dataset. Analysis of the data generates a descriptive format in the form of decision trees. The decision tree checks an attribute at each node and the decision is made to classify the data. They are easy to interpret thus the decision trees of various species of plant are compared and the relevance of attribute is calculated.
• **Performance Evaluation tables**
The predictive performance was calculated by WEKA software. The TP rates, FP rates, precision (specificity) and recall (sensitivity) values. The values which were near the value one were considered good for classification. F-measure is the harmonic mean of the precision and recall. It is the threshold of precision and recall as they both cannot be increased together.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>TP Rate</th>
<th>FP Rate</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training set</td>
<td>0.95</td>
<td>0.015</td>
<td>0.912</td>
<td>0.913</td>
<td>0.966</td>
<td>T</td>
</tr>
<tr>
<td>Cross-validation with fold 10</td>
<td>0.98</td>
<td>0.049</td>
<td>0.912</td>
<td>0.912</td>
<td>0.957</td>
<td>F</td>
</tr>
<tr>
<td>Cross-validation with fold 5</td>
<td>0.94</td>
<td>0.065</td>
<td>0.915</td>
<td>0.94</td>
<td>0.938</td>
<td>T</td>
</tr>
<tr>
<td>Percentage split (66%)</td>
<td>0.95</td>
<td>0.024</td>
<td>0.915</td>
<td>0.915</td>
<td>0.937</td>
<td>F</td>
</tr>
<tr>
<td>Test against Arabidopsis thaliana</td>
<td>0.86</td>
<td>0.06</td>
<td>0.869</td>
<td>0.91</td>
<td>0.901</td>
<td>T</td>
</tr>
<tr>
<td>Test against Arabidopsis thaliana</td>
<td>0.94</td>
<td>0.131</td>
<td>0.975</td>
<td>0.94</td>
<td>0.908</td>
<td>F</td>
</tr>
<tr>
<td>Test against Sorghum bicolor</td>
<td>0.95</td>
<td>0.038</td>
<td>0.91</td>
<td>0.933</td>
<td>0.931</td>
<td>T</td>
</tr>
<tr>
<td>Test against Sorghum bicolor</td>
<td>0.91</td>
<td>0.047</td>
<td>0.914</td>
<td>0.932</td>
<td>0.932</td>
<td>F</td>
</tr>
<tr>
<td>Test against Zea mays</td>
<td>0.92</td>
<td>0.056</td>
<td>0.936</td>
<td>0.892</td>
<td>0.913</td>
<td>T</td>
</tr>
<tr>
<td>Test against Zea mays</td>
<td>0.93</td>
<td>0.108</td>
<td>0.927</td>
<td>0.939</td>
<td>0.917</td>
<td>F</td>
</tr>
<tr>
<td>Test against Oryza sativa</td>
<td>0.95</td>
<td>0.065</td>
<td>0.916</td>
<td>0.957</td>
<td>0.946</td>
<td>T</td>
</tr>
<tr>
<td>Test against Oryza sativa</td>
<td>0.95</td>
<td>0.043</td>
<td>0.956</td>
<td>0.935</td>
<td>0.945</td>
<td>F</td>
</tr>
</tbody>
</table>

**Table 3:** Classification results with reference to *Oryza sativa* (9)

5. **CONCLUSION**
The values for true positives and false positives were found to be significant. Having obtained the classification results based on the decision trees we conclude that this approach is best suited for the classification of miRNAs. In our studies we predicted the dominating attributes such as DFL, BPN, GCC, AUC etc. These attributes are the basis of classification of miRNAs of related species of plants. These attributes are biologically significant. The decision trees obtained from the datasets of different species will be useful to study and classify our data in a better way.

6. **FUTURE SCOPE**
In future we plan to classify larger datasets from various other species. We also plan to make a web enabled tool for prediction of miRNA based on the dominating attributes. We can also use
Feature based Analysis and Classification of Plant MicroRNAs

other data mining tools to explore more possibilities of miRNA classification.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>TP Rate</th>
<th>FP Rate</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training set</td>
<td>0.95</td>
<td>0.04</td>
<td>1.00</td>
<td>0.95</td>
<td>0.97</td>
<td>T</td>
</tr>
<tr>
<td>Cross-validation with fold 10</td>
<td>0.925</td>
<td>0.06</td>
<td>0.948</td>
<td>0.925</td>
<td>0.935</td>
<td>T</td>
</tr>
<tr>
<td>Percentage split (66%)</td>
<td>0.924</td>
<td>0.06</td>
<td>0.936</td>
<td>0.924</td>
<td>0.938</td>
<td>F</td>
</tr>
<tr>
<td>Test against <em>Arabidopsis thaliana</em></td>
<td>0.949</td>
<td>0.103</td>
<td>0.9</td>
<td>0.949</td>
<td>0.924</td>
<td>T</td>
</tr>
<tr>
<td>Test against <em>Zea mays</em></td>
<td>0.884</td>
<td>0.651</td>
<td>0.904</td>
<td>0.898</td>
<td>0.92</td>
<td>F</td>
</tr>
<tr>
<td>Test against <em>Sorghum bicolor</em></td>
<td>0.924</td>
<td>0.189</td>
<td>0.931</td>
<td>0.924</td>
<td>0.875</td>
<td>T</td>
</tr>
<tr>
<td>Test against <em>Brassica napus</em></td>
<td>0.812</td>
<td>0.766</td>
<td>0.894</td>
<td>0.812</td>
<td>0.86</td>
<td>F</td>
</tr>
<tr>
<td>Test against <em>Sorghum bicolor</em></td>
<td>0.835</td>
<td>0.101</td>
<td>0.897</td>
<td>0.835</td>
<td>0.891</td>
<td>F</td>
</tr>
<tr>
<td>Test against <em>Brassica napus</em></td>
<td>0.899</td>
<td>0.115</td>
<td>0.887</td>
<td>0.899</td>
<td>0.893</td>
<td>F</td>
</tr>
</tbody>
</table>

Table 4: Classification results with reference to *Arabidopsis thaliana* (14)

<table>
<thead>
<tr>
<th>Dataset</th>
<th>TP Rate</th>
<th>FP Rate</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training set</td>
<td>0.971</td>
<td>0.071</td>
<td>1.00</td>
<td>0.971</td>
<td>0.985</td>
<td>T</td>
</tr>
<tr>
<td>Cross-validation with fold 10</td>
<td>0.924</td>
<td>0.033</td>
<td>0.948</td>
<td>0.924</td>
<td>0.915</td>
<td>T</td>
</tr>
<tr>
<td>Percentage split (66%)</td>
<td>0.959</td>
<td>0.03</td>
<td>0.945</td>
<td>0.959</td>
<td>0.929</td>
<td>T</td>
</tr>
<tr>
<td>Test against <em>Arabidopsis thaliana</em></td>
<td>0.919</td>
<td>0.085</td>
<td>0.915</td>
<td>0.919</td>
<td>0.917</td>
<td>T</td>
</tr>
<tr>
<td>Test against <em>Sorghum bicolor</em></td>
<td>0.939</td>
<td>0.045</td>
<td>0.949</td>
<td>0.939</td>
<td>0.931</td>
<td>F</td>
</tr>
<tr>
<td>Test against <em>Brassica napus</em></td>
<td>0.925</td>
<td>0.181</td>
<td>0.855</td>
<td>0.925</td>
<td>0.893</td>
<td>F</td>
</tr>
<tr>
<td>Test against <em>Sorghum bicolor</em></td>
<td>0.946</td>
<td>0.203</td>
<td>0.824</td>
<td>0.946</td>
<td>0.811</td>
<td>T</td>
</tr>
<tr>
<td>Test against <em>Brassica napus</em></td>
<td>0.707</td>
<td>0.034</td>
<td>0.957</td>
<td>0.707</td>
<td>0.811</td>
<td>F</td>
</tr>
<tr>
<td>Test against <em>Brassica napus</em></td>
<td>0.913</td>
<td>0.283</td>
<td>0.764</td>
<td>0.913</td>
<td>0.812</td>
<td>T</td>
</tr>
<tr>
<td>Test against <em>Brassica napus</em></td>
<td>0.717</td>
<td>0.067</td>
<td>0.922</td>
<td>0.717</td>
<td>0.795</td>
<td>F</td>
</tr>
</tbody>
</table>

Table 5: Classification results with reference to *Zea mays* (20)

REFERENCES

[15] X. Wang, J. Zhang, F. Li, J. Gu, T. He, X. Zhang and Y. Li, "MicroRNA identification based on sequence and...


