Analysis and Classification of Plant MicroRNAs Using Decision Tree Based Approach

A.K. Mishra and H. Chandrasekharan

AKMU, Indian Agricultural Research Institute, New Delhi

Abstract. MicroRNA (miRNA) analysis have progressed tremendously in recent past, but further indepth computational study is required to know the complete potential of these RNAs. Due to its short length (~20 nucleotides), it is difficult to use the conventional lab techniques for microRNA prediction and analysis. This has led to this work in the domain of computational biology. These are the non coding small RNAs which are responsible for the gene regulation at the post translational level by binding to the mRNAs and thereby stopping the translation activities. Therefore,the effect of microRNAs on the various proteins is important. In this paper we have studied 1010 microRNA and precursor microRNA sequences from monocots . Our study in this paper is on the microRNA classification using decision trees and determining dominating attributes. We have used WEKA, a data mining tool which helps us to study the large data and classify it. The decision trees based classification was best suited for the miRNA study and the dominating attributes derived are biologically significant.

Keywords: miRBase, precursor, RNAFold, WEKA, J48, Decision Trees.

1 Introduction

MicroRNAs 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, microRNA 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 microRNA (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 microRNA. 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), on the basis of the stability of the 5'end. In plants as Drosha is lacking so Dicer performs the processing [1, 2].

Mature miRNA are partially complimentary to messenger RNA and they play an important role in gene regulation through mRNA cleavage or translational repression by associating with the RISC. Prediction of miRNA helps us to understand its structure and thus its function and role in organism. miRNA function in cell death,

S.C. Satapathy et al. (eds.), *ICT and Critical Infrastructure: Proceedings of the 48th Annual Convention of CSI - Volume II*, Advances in Intelligent Systems and Computing 249,
DOI: 10.1007/978-3-319-03095-1_12, © Springer International Publishing Switzerland 2014

proliferation and fat metabolism in *Drosophila melanogaster*[3]. 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 organisms.Need for computational prediction: The short length of microRNA makes it difficult to analyse it with the help of conventional genetic techniques. Some microRNAs 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 [4]. 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 [5].

2 Review of Literature

To carry out the computational prediction and their analysis there are some tools which are based on the following techniques.

- Filter based- this approach uses different features and conservation criteria to restrict the presursor candidates.
- 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.

3 Materials and Methods

3.1 Reference miRNAs

The set of miRNAs and precursor miRNAs we used were downloaded from miRBase (version15, http://www.mirbase.org/). 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, Arabidopsis thaliana are more in number as their genome sequence information is available.

Table 1. Tools for computational prediction of miRNA sequences are:

Name	Type	URL	Techniques
Mir Scan	W	http://genes.mit.edu/mirscan/	Filter-based
MiRFinder	D	http://www.bioinformatics.org/mirfinder/	Filter-based
ProMIR	W	http://cbit.snu.ac.kr/_ProMiR2/	Machine learning
TripletSVM	D	http://bioinfo.au.tsinghua.edu.cn/mirnasvm/	Machine learning
RNAMicro	D	http://www.bioinf.uni-leipzig.de/_jana/software/RNAmicro.html	Machine learning
MiPred	W	http://www.bioinf.seu.edu.cn/miRNA/	Machine learning
Mireval	W	http://tagc.univ-mrs.fr/mireval/	Mixed approaches
findMiRNA	D	http://sundarlab.ucdavis.edu/mirna/download.html	Target based
MirAlign	W	http://bioinfo.au.tsinghua.edu.cn/miralign/	Homology-based
BayesmiRNAfind	W	http://wotan.wistar.upenn.edu/miRNA	Rule based

3.2 Preparation of Dataset

Using the PERL scripts we automated the retrieval of 1010 sequences from miRBase. These sequences were put into the RNAfold, a software developed by M. Zuker and P. Stiegler [6, 7, 8, 9] for the secondary structure of our sequences and their MFE. Coding in PERL was done to calculate the values of the set attributes from their secondary structures.

3.3 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 is the easiest and simplest way to interpret the results.

3.4 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.

>osa-MIR395s MI0001037
GUAUCACCGUGAGUUCCCUUCAAGCACUUCACGUGGCACUAUUUCAAU
GCCUAUU GUGAAGUGUUUGGGGGAACUCUCGAUGUUCC

>osa-miR395s MIMAT0000968
GUGAAGUGUUUGGGGGAACUC

3.5 Sequences from miRBase and Their Ids

>osa-MIR395s MI0001037

GUAUCACCGUGAGUUCCCUUCAAGCACUUCACGUGGCACUAUUUCAAU
GCCUAUU GUGAAGUGUUUGGGGGAACUCUCGAUGUUCC

....((.((.(((((((((.(((((((((((((.((((.........))))...)))))))))))))))))))))).)).))....

RNA fold dot bracket secondary structure

3.6 Identification of Attributes and Calculating the Values

From the sequences, 9 and 14 attributes were considered for mature microRNA sequences and precursor sequences 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 Uracil nucleotide content in the sequence, MSK maximum stack 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 [10].

Table 2. Attributes for mature/precursor microRNA

Attributes for mature microRNA	Attributes for precursor microRNA
ARM, DFL, BPN, LNM, POP, GCC, MFE, DAS, DAE	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 sets of datasets were prepared for different species.

3.7 Attributes

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

3.8 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 building of the plant species using a classifier.

The graphical view represents some patterns found in the microRNA sequences. In the sequences the dominating nucleotide at the beginning of the sequence 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).

4.1 Decision Tree Construction

The datasets were fed into the software and run. A graphical representation of the dataset was displayed. This graphical view indicates various patterns in out 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 3. Selected attributes

	Oryza sativa (20)	*Arabidopsis thaliana* (20)	*Zea mays*(20)	*Sorghum bicolor*(20)	*Brassica napus*((20)
BestFirst+CfsSubset Ev al	BLR, NHP, FRE, FEN, MSK, NTL, DFL	NBP, BLR, NHP, FEN, MTL, DFL	BLR, NHP, FRE, FEN, SBR, DFL, DAE	BLR, FRE, FEN, MSK, SDI, MBA, MBS,	BLR, FEN, SBR, MBA, MBS, DFL
Ranker+ChiSquared At tributeEval	FEN, NHP, FRE, DFL, SDI, MSK	FEN, BLR, DFL, NHP, FRE, MTL, MSK, SDI	BLR, FEN, DFL, NHP, MSK, FRE, SDI, MTL	MBS, BLR, FEN, SDI, MBA, DFL, FRE	MBS, FEN, BLR, SDI, MBA, FRE, NHP, DFL
GreedyStepwise+Co ns istencySubsetEval	BLR, NHP, FRE, FEN, MSK, SDI, MBS, DFL	NBP, BLR, NHP, FRE, FEN, MSK	BLR, NHP, FRE, FEN, MSK, MBL, BS, MTL, DFL	BLR, MBS	BLR, MBS
Ranker+ SVMAttribute Eval	NHP, DFL, FEN, AUC, FRE,	FEN, BLR, FRE, NHP, AUC, MBS	BLR, NHP, DFL, FEN, AUC, FRE, SDI	MBS, MBA, BLR, MSK, SDI, FEN, AUC	FEN, BLR, MBS, MBA, SDI, AUC, DFL, NHP
Ranker+InfoGainAt tri buteEval	NHP, DFL, FEN, AUC, FRE, MBS, MTL, SDI	FEN, BLR, DFL, NHP, FRE, MTL, MSK	BLR, FEN, DFL, NHP, MSK, FRE, SDI,MBL	MBS, BLR, FEN, SDI, MBA, DFL, FRE	MBS, BLR, FEN, MBA, SDI, DFL, NHP

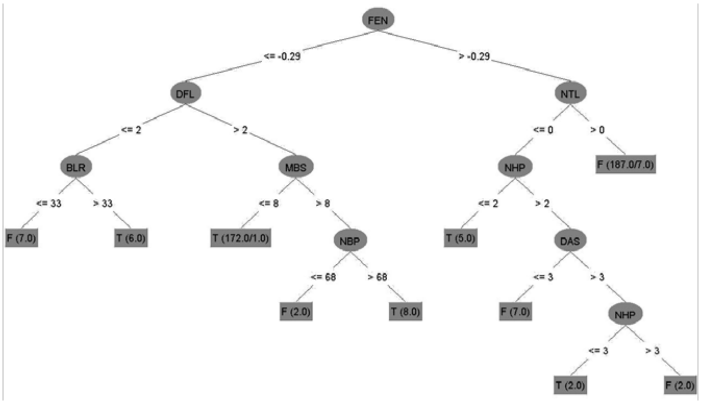

Fig. 1. Decision tree for with 20 attributes *Arabidopsis thaliana*

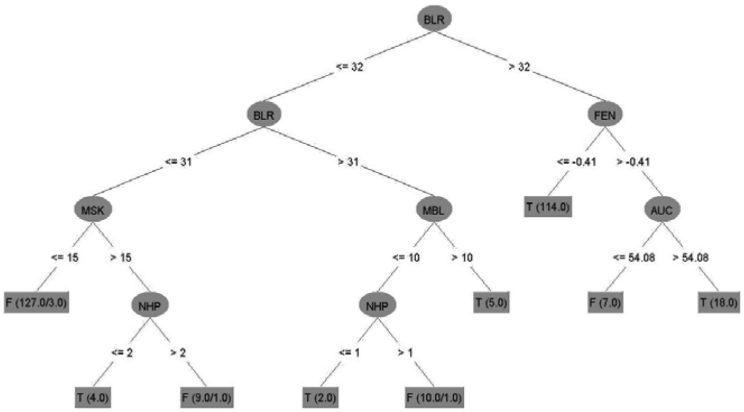

Fig. 2. Decision tree with 14 attributes for *Sorghum bicolor*

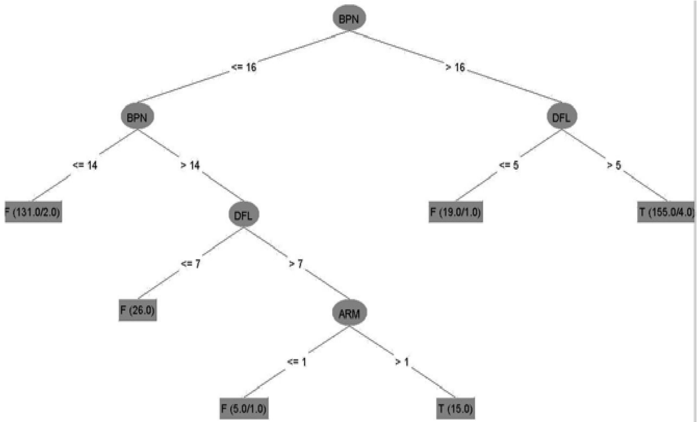

Fig. 3. Decision tree with 9 attributes for *Zea mays*

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.

4.2 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 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 4. Classification results with reference to *Oryza sativa*(9)

	TP Rate	FP Rate	Precision	Recall	F-Measure	Class
Training set	0.951	0.018	0.982	0.951	0.966	T
	0.982	0.049	0.952	0.982	0.967	F
Cross-validation with fold 10	0.94	0.065	0.935	0.94	0.938	T
	0.935	0.06	0.939	0.935	0.937	F
Percentage split (66%)	0.956	0.024	0.97	0.956	0.963	T
	0.976	0.044	0.965	0.976	0.971	F
Test against *Arabidopsis thaliana*	0.869	0.06	0.935	0.869	0.901	T
	0.94	0.131	0.878	0.94	0.908	F
Test against *Zea mays*	0.953	0.088	0.91	0.953	0.931	T
	0.912	0.047	0.954	0.912	0.932	F
Test against *Sorghum bicolor*	0.892	0.061	0.936	0.892	0.913	T
	0.939	0.108	0.897	0.939	0.917	F
Test against *Brassica napus*	0.957	0.065	0.936	0.957	0.946	T
	0.935	0.043	0.956	0.935	0.945	F

Table 5. Classification results with reference to *Arabidopsis thaliana* (14)

	TP Rate	FP Rate	Precision	Recall	F-Measure	Class
Training set	0.96	0	1	0.96	0.979	T
	1	0.04	0.961	1	0.98	F
Cross-validation with fold 10	0.925	0.05	0.948	0.925	0.936	T
	0.95	0.075	0.926	0.95	0.938	F
Percentage split (66%)	0.924	0	1	0.924	0.961	T
	1	0.076	0.903	1	0.949	F
Test against *Oryza thaliana*	0.949	0.105	0.9	0.949	0.924	T
	0.895	0.051	0.946	0.895	0.92	F
Test against *Zea mays*	0.924	0.188	0.831	0.924	0.875	T
	0.812	0.076	0.914	0.812	0.86	F
Test against *Sorghum bicolor*	0.885	0.101	0.897	0.885	0.891	T
	0.899	0.115	0.887	0.899	0.893	F
Test against *Brassica napus*	0.957	0.043	0.957	0.957	0.957	T
	0.957	0.043	0.957	0.957	0.957	F

Table 6. Classification results with reference to *Zea mays*(20)

	TP Rate	FP Rate	Precision	Recall	F-Measure	Class
Training set	0.971	0	1	0.971	0.985	T
	1	0.029	0.971	1	0.986	F
Cross-validation with fold 10	0.924	0.053	0.946	0.924	0.935	T
	0.947	0.076	0.925	0.947	0.936	F
Percentage split (66%)	0.929	0.05	0.945	0.929	0.937	T
	0.95	0.071	0.934	0.95	0.942	F
Test against *Oryza sativa*	0.919	0.085	0.915	0.919	0.917	T
	0.915	0.081	0.919	0.915	0.917	F
Test against *Arabidopsis thaliana*	0.839	0.045	0.949	0.839	0.891	T
	0.955	0.161	0.856	0.955	0.903	F
Test against *Sorghum bicolor*	0.946	0.203	0.824	0.946	0.881	T
	0.797	0.054	0.937	0.797	0.861	F
Test against *Brassica napus*	0.913	0.283	0.764	0.913	0.832	T
	0.717	0.087	0.892	0.717	0.795	F

5 Conclusion

The classification yielded good results based on the decision trees which are best suited for the classification of miRNAs. In our studies we predicted the dominating attributes which are the basis o f classification of miRNAs of related species. These attributes hold biological significance. Clustering was not required as our data was based on two classes.

References

1. He, L., Hannon, G.J.: MicroRNAs: small RNAs with a big role in gene regulation. Nature Genetics (2004)
2. Lee, Y., Jeon, K., Lee, J.-T., Kim, S., Kim, V.N.: MicroRNA maturation: stepwise processing and subcellular localization. EMBO J. (2002)
3. Alvarez-Garcia, I., Miska, E.A.: MicroRNA functions in animal development and human disease. Development. The Company of Biologists (2005)
4. Mendes, N.D., Freitas, A.T., Sagot, M.-F.: Current tools for the identification of miRNA genes and their targets. Nucleic Acids Research (May 2009)
5. Wang, X., Zhang, J., Li, F., Gu, J., He, T., Zhang, X., Li, Y.: MicroRNA identification based on sequence and structure alignment. Bioinformatics (2005)
6. Hofacker, I.L., Fontana, W., Stadler, P.F., Bonhoeffer, S., Tacker, M., Schuster, P.: Fast-Folding and Comparison of RNA Secondary Structures. Monatshefte F. Chemie 125, 167–188 (1994)

7. Zuker, M., Stiegler, P.: Optimal computer folding of large RNA sequences using thermodynamic and auxiliary information. Nucl. Acid Res. (1981)
8. Hofacker, I.L., Stadler, P.F.: Memory Efficient Folding Algorithms for Circular RNA Secondary Structures. Bioinformatics (2006)
9. Bompfunewerer, A.F., Backofen, R., Bernhart, S.H., Hertel, J., Hofacker, I.L., Stadler, P.F., Will, S.: Variations on Folding and Alignment: Lessons from Benasque. J. Math. Biol. (2007)
10. Mishra, A.K., Lobiyal, D.K.: Exploring Dominating Features from Apis Mellifera Pre-miRNA. IEEE (2009)