PRESENT STATUS OF IN SILICO APPROACH FOR MICRO-RNA IDENTIFICATION

T. Komathy[1], M.I.S. Safeena[2]& M.C.M. Zakeel[2]

[1,2]Department of Biological Sciences, Faculty of Applied Sciences, South Eastern University of Sri Lanka, Sammanthurai, Sri Lanka

[2]Department of Plant Sciences, Faculty of Agriculture, Rajarata University of Sri Lanka, Puliyankulama, Anuradhapura, Sri Lanka

Komss01@gmail.com, safeenim@seu.ac.lk, zakeelag48@yahoo.com

ABSTRACT: MicroRNAs (miRNAs) are endogenous, small, noncoding RNAs of 18-25 nucleotides in length that negatively regulate their complementary messenger-RNAs at the post transcriptional level. Three methods are principally deployed to identify miRNA such as classic cloning, sequencing and computational approach. Computational approach is carried out mainly based on bioinformatics which is using small RNA library construction and sequencing to find miRNAs in the known genotypes (sequences). Another technique called the classical approach that is an example of forward genetics in which researchers determine the unknown genotype (miRNA sequence) of a known phenotype. The miRNAs have shown both evolutionarily converged nature (conserved miRNAs) from species to species within the same kingdom and species specific expression (species specific miRNAs). Majority of the miRNA genes are observed as conserved miRNAs as orthologs. This conserved nature of majority of miRNAs becomes an important logical tool for bioinformatics discovery of miRNAs in other species. Identification of miRNAs by using bioinformatics tools is now a commonplace and of the most widely used methods and it has facilitated the prediction of new miRNAs in both plant and animal systems. This is largely used due to its low cost and high efficiency.

Keywords: miRNA, EST, Bioinformatics, Computational Methods.

1. INTRODUCTION

MicroRNAs (miRNAs) are a class of small (~21nt) non-coding RNA molecules found in animals, plants, and some viruses and play pivotal roles in gene expression at transcriptional and post-transcriptional levels. These RNAs are evolutionarily conserved across species, functioning as central component in a wide range of biological processes such as metabolism, development, host-pathogen interaction, and disease. In plants, miRNAs function to control tissue differentiation and development, signal transduction, vegetative or reproductive growth and the response to biotic and abiotic stresses such as drought, salinity and pathogens (Zhang et al., 2006). Therefore, identification of miRNAs would be highly useful to interfere with gene expression in all organisms to enhance production and control diseases. There is a potential to use miRNAs as therapeutics in human particularly to treat diseases like cancer.

Three methods are principally deployed to identify miRNA: (1) classic cloning method, (2) sequencing method, and (3) computational approach. Computational approach can further be divided into three types: (i) *ab initio* prediction based on the sequence and structural features, (ii) comparative genomic strategy based on evolutionary conservation, and (iii) integrated approach (Padmashree and Ramachandraswamy, 2015).

MiRNAs were initially identified by a genetic screening technology (Lee et al., 1993; Wightman et al., 1993). Although this approach is much useful for identifying miRNAs, it is highly limited due to the expense, time-consuming nature, and its domination by chance (Lai et al., 2003 and Zhang et al., 2006). Therefore, the present trend in miRNA identification is via *in silico* approach and at present, many miRNAs have been identified in various plants of horticultural and medicinal importance via this approach.

Having understood the importance of miRNAs and their identification through *in- silico* methods, we have in this paper reviewed various computational methods used to identify miRNAs and the similarities and differences among these methods. In addition, we have reviewed the prediction tools for the potential targets of the identified miRNAs.

2. OVERVIEW OF MICRO-RNA PREDICTION

Various *in silico* (computational) approaches have been developed for the successful identification of miRNAs in various plants and animals, including human, *Caenorhabditis elegans*, *Camellia sinensis*, *Arabidopsis thaliana*, and *Oryza sativa*. However, all these methods have many steps in common though a few steps differ among each other. Reviewing all these methods would be of immense use to assess their efficacy in identifying miRNAs using ESTs (Expressed Sequence Tags) and GSSs (Genomic Survey Sequences) and this has not been done yet making it a timely need. Almost all the methods reviewed here follow a more or less common procedure for the identification of miRNAs as shown in figure 1.

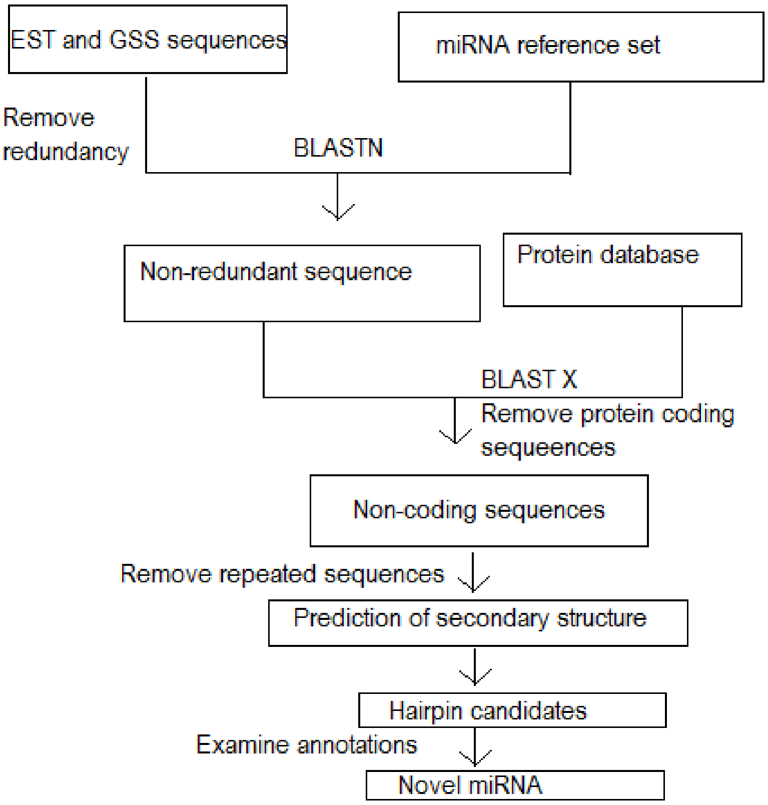


Figure 1. General flow diagram for the prediction of novel potential miRNA

2.1. *Retrieval of EST, GSS and Available miRNA Sequences*

NCBI is the commonly used database to collect ESTs and GSSs (Zhang et al., 2006). To predict miRNAs, homology searches against available miRNAs are performed. Though many databases are available for retrieval of miRNAs, the database called miRBase is predominantly used by many researchers (Zhang et al., 2006).

2.2. BLAST

BLAST (Basic Local Alignment Search Tool) is a method used to find regions of local similarity among nucleotide or amino acid sequences. It compares a query sequence (DNA or protein) to a large set of sequences (the target) and calculates the statistical significance of matches. ENSEMBL, from its release 71 onwards, uses the NCBI Blast for its search options. Comparing DNA query with DNA database is known as BLASTN whereas the comparison with protein database is BLASTX.

Prediction of potential miRNAs is a tedious and cautious procedure. With the advent of novel and powerful informatics infrastructure as well as bioinformatics tools, the possibility to discover novel microRNA and interactions in complex datasets became feasible. Recent studies have shed some light over the function of miRNA ,which may lead to numerous potential applications from infectious diseases control, cancer development decrease, and inhibition of protein synthesis to improvement of plant production in agribusiness(as revised by Pillai,2005). The reference sequences are used as a query for homology search against local nucleotide sequence database at e-value threshold < 0.01 using ncbiblast + 2.2.28 program (Table. 1). Removal of above protein coding sequences results the output of BLASTX as shown in the Table 2.

Query Id	Subject Id	% Identity	Alignment Length	Mismatches	Gap Opens	q. start	q. end	s. start	s. end	evalue	Bit Score	Query Lenth	Subject Length
sme-miR-2180	gi|198407369|gb|GD178969.1|GD178969	100	17	0	0	4	20	359	375	0.004	32.5	20	469
gma-miR4993	gi|327340495|gb|JG704482.1|JG704482	100	17	0	0	3	19	455	471	0.005	32.5	21	770
ppc-miR-8282c-5p	gi|198411421|gb|GD175540.1|GD175540	100	17	0	0	1	17	80	64	0.005	32.5	22	401
zma-miR397b-3p	gi|198408671|gb|GD177933.1|GD177933	95	20	1	0	2	21	606	587	0.005	32.5	21	695
tae-miR9773	gi|198407337|gb|GD179399.1|GD179399	96	23	0	1	1	23	217	238	5.00E-04	36.2	24	568
osa-miR414	gi|327339103|gb|JG703090.1|JG703090	100	20	0	0	1	20	221	240	1.00E-04	38.1	21	323
osa-miR414	gi|327340098|gb|JG704085.1|JG704085	100	20	0	0	1	20	221	240	1.00E-04	38.1	21	323
oan-miR-1421t-5p	gi|198409122|gb|GD175431.1|GD175431	100	17	0	0	2	18	277	293	0.005	32.5	22	539
atr-miR8587	gi|198407132|gb|GD176058.1|GD176058	100	17	0	0	1	17	135	119	0.007	32.5	24	591
ppy-miR-1255a	gi|4521717|gb|AI563335.1|AI563335	95	20	1	0	1	20	183	202	0.006	32.5	23	296
ath-miR414	gi|327339103|gb|JG703090.1|JG703090	95	21	1	0	1	21	221	241	0.001	34.4	21	323
ath-miR414	gi|327340098|gb|JG704085.1|JG704085	95	21	1	0	1	21	221	241	0.001	34.4	21	323
ath-miR414	gi|4521755|gb|AI563373.1|AI563373	100	17	0	0	1	17	119	135	0.005	32.5	21	382

Table 1. Results of a BLASTN search

Query Id	Subject Id	% Identity	Alignment Length	Mismatches	Gap Opens	q. start	q. end	s. start	s. end	evalue	Bit Score
gi|327339301|gb|JG703288.1|JG703288	gi|778698418|ref|XP_011654530.1|	94.98	279	14	0	3	839	174	452	0	555
gi|327339301|gb|JG703288.1|JG703288	gi|659072700|ref|XP_008466806.1|	94.22	277	16	0	3	833	174	450	0	548
gi|327339301|gb|JG703288.1|JG703288	gi|763747867|gb|KJB15306.1|	89.89	277	28	0	3	833	79	355	0	527
gi|327339301|gb|JG703288.1|JG703288	gi|823134732|ref|XP_012467166.1|	89.89	277	28	0	3	833	174	450	0	528
gi|327339301|gb|JG703288.1|JG703288	gi|1028963869|ref|XP_016725437.1|	89.89	277	28	0	3	833	174	450	0	528
gi|327339301|gb|JG703288.1|JG703288	gi|728841645|gb|KHG21088.1|	89.89	277	28	0	3	833	174	450	0	528
gi|327339301|gb|JG703288.1|JG703288	gi|567911473|ref|XP_006448050.1|	89.53	277	29	0	3	833	44	320	0	522
gi|327339301|gb|JG703288.1|JG703288	gi|590697286|ref|XP_007045397.1|	90.25	277	27	0	3	833	174	450	0	526
gi|327339301|gb|JG703288.1|JG703288	gi|763783339|gb|KJB50410.1|	89.89	277	28	0	3	833	79	355	0	523
gi|327339301|gb|JG703288.1|JG703288	gi|703133968|ref|XP_010105520.1|	89.61	279	29	0	3	839	174	452	0	526
gi|327339301|gb|JG703288.1|JG703288	gi|641821493|gb|KDO41165.1|	89.89	277	28	0	3	833	77	353	0	522
gi|327339301|gb|JG703288.1|JG703288	gi|1029085808|ref|XP_016703107.1|	89.53	277	29	0	3	833	174	450	0	525
gi|327339301|gb|JG703288.1|JG703288	gi|728833984|gb|KHG13427.1|	89.53	277	29	0	3	833	174	450	0	525
gi|327339301|gb|JG703288.1|JG703288	gi|572152989|ref|NP_001275838.1|	89.53	277	29	0	3	833	174	450	0	524
gi|327339301|gb|JG703288.1|JG703288	gi|694382255|ref|XP_009367152.1|	88.61	281	32	0	3	845	174	454	0	524
gi|327339301|gb|JG703288.1|JG703288	gi|823210917|ref|XP_012438388.1|	89.89	277	28	0	3	833	174	450	0	523
gi|327339301|gb|JG703288.1|JG703288	gi|778698422|ref|XP_004151335.2|	90.84	273	25	0	3	821	174	446	0	523
gi|327339301|gb|JG703288.1|JG703288	gi|694382196|ref|XP_009367130.1|	88.26	281	33	0	3	845	174	454	0	522
gi|327339301|gb|JG703288.1|JG703288	gi|658045009|ref|XP_008358174.1|	88.26	281	33	0	3	845	44	324	0	516

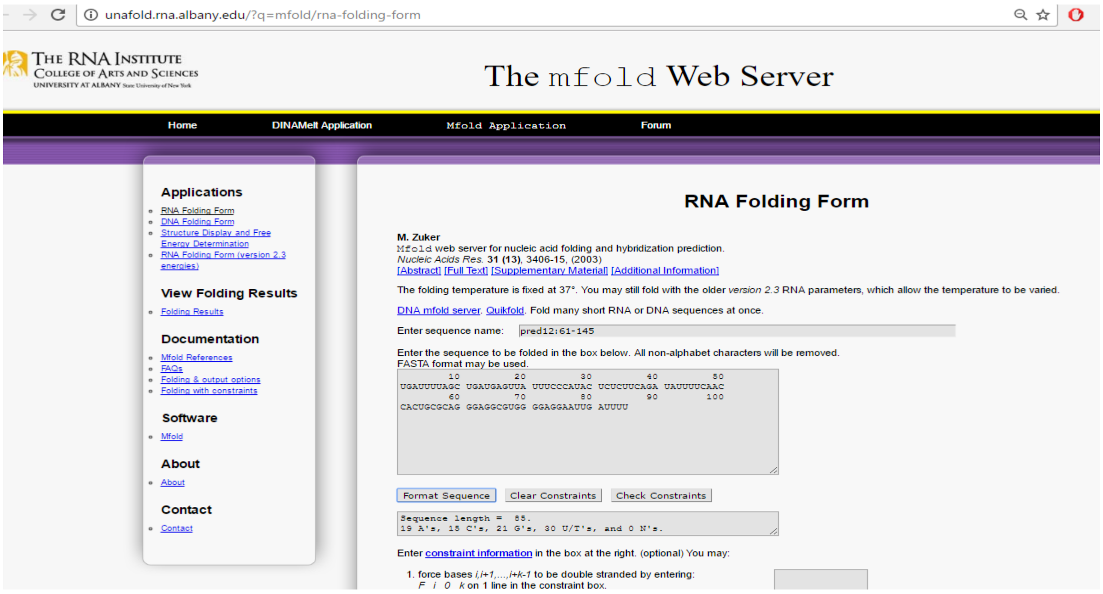
Table 2. Results obtained after a BLASTX search

2.3. Secondary Structure Prediction

miRNAs assume specific hairpin secondary structures before being exported to the cytoplasm. Though several bioinformatics tools such as PMRD, microPC and Mfold are used to predict the secondary structures of pre-miRNA sequences, Mfold software developed by Zuker (2003) is the most commonly used software tool by many researchers.

Precursor sequences of potential miRNA homolog are used for secondary structure prediction using the Zuker folding algorithm with MFOLD 3.1, which is publicly available at www.bioinfo.rpi.edu/applications/mfold/old/rna/. The following parameters are used in predicting the secondary structures: (1) Linear RNA sequence, (2) it contains ~22 nt mature miRNA sequence within one arm of the hairpin, (3) an MFEI of greater than 0.8536, (4) 30–70% A+U content, (5) predicted mature miRNAs has no more than six mismatches with the opposite miRNA* sequence in the other arm, (6) maximum size of 3 nt for a bulge in the miRNA sequence, and (7) No loop or break in miRNA sequences is allowed. These criteria significantly reduce false positives. ΔG values (kcal/mol) of stem-loop structures generated by MFOLD program are used to calculate their negative minimal free energies (MFEs), which is directly correlated with the sequence length. To normalize the potential effect of sequence length on MFE and to differentiate miRNAs from other RNAs, two energy measurements namely adjusted minimal folding energy (AMFE) and minimal folding free energy index (MFEI) are used (Zuker, 2003). AMFE is the MFE of a 100 nucleotide sequence.

It is important to note here that retrieval of sequence (ESTs or GSSs), BLASTing and secondary structure prediction are common among all the approached used to predict miRNAs computationally. However, the subsequent steps such as removal of redundant sequences and the prediction of precursor sequences differ among the approaches. Precursor sequences and respective hairpin structure predictions are obtain through mirEvl in addition to Zuker folding algorithm with Mfold-3.1 (Fig. 2). MirEval is the first miRNA search tool that allows a thorough multi-criteria analysis of input sequences and delivers an unbiased, clear report. One of our main concerns was to ensure that, this tool is easy to maintain, long lasting and impervious to version changes (Gao et al., 2013).



RNA Folding Form

M. Zuker
Mfold web server for nucleic acid folding and hybridization prediction.
Nucleic Acids Res. **31 (13)**, 3406-15, (2003)
[Abstract] [Full Text] [Supplementary Material] [Additional Information]

The folding temperature is fixed at 37°. You may still fold with the older *version 2.3* RNA parameters, which allow the temperature to be varied.

DNA mfold server. Quikfold. Fold many short RNA or DNA sequences at once.

Enter sequence name: pred12:61-145

Enter the sequence to be folded in the box below. All non-alphabet characters will be removed.
FASTA format may be used.

```
        10         20         30         40         50
UGAUUUUAGC UGAUGAGUUA UUUCCCAUAC UCUCUUCAGA UAUUUUCAAC
        60         70         80         90        100
CACUGCGCAG GGAGGCGUGG GGAGGAAUUG AUUUU
```

```
Sequence length =  85.
19 A's, 15 C's, 21 G's, 30 U/T's, and 0 N's.
```

Enter constraint information in the box at the right. (optional) You may:

1. force bases *i,i+1,...,i+k-1* to be double stranded by entering:
   F i 0 k on 1 line in the constraint box.

Figure 2. Intermediate step in prediction of potential miRNAs through MFOLD

2.4. Removal of Redundancy

Duplicates in the reference set of miRNAs are removed by performing multiple sequence alignment using ClustalW to avoid redundancy (Dehury et al., 2013). The redundancies in the EST sequences are removed by performing sequence assembly using locally installed CAP3 program with the default parameters (Dehury et al., 2013). However, some people use a sequence assembly program, EGassembler (http://www.genome.jp/tools/egassembler/), to remove redundancy. Repeated sequences of miRNAs are removed with the Jalview program with the threshold value of 100. For the precursor prediction MirEval's report is condensed be an easy to read, color-coded output. It is a comprehensive tool, easy to use and very informative.

It will allow users with no prior knowledge of *in-silico* detection of microRNAs to take advantage of the most successful approaches to investigate sequences of interest. The biopython language is also nowadays widely used by researchers to remove redundancy.

2.5. Precursor Prediction

After BLASTX, the reference sequences are subjected to precursor prediction. miREval 2.0 is a commonly used software for precursor prediction(Dehury et al., 2013). Nevertheless, there are many software tools available for this purpose. The most commonly used RNA secondary structure prediction tools is mfold. It provides many facilities such as highlighting the mature sequence in a pre-miRNA. The predicted secondary structure can be viewed in various formats such as pdf, png, jpg etc (Fig. 3). It also predicts the thermodynamic details for folded structures e.g. minimum folding energy (mfe). Besides this mfold allows the user to store and draw the predicted structure. The following criteria were applied in designating the RNA sequence as described by Wang et al., 2004.

- No more than 6 mismatches are between the predicted mature miRNA sequence and its opposite miRNA.
- (miRNA*) sequence in the secondary structure.
- No loop or break is in the miRNA or miRNA* sequences.
- Predicted secondary structure has higher MFEI and negative MFE

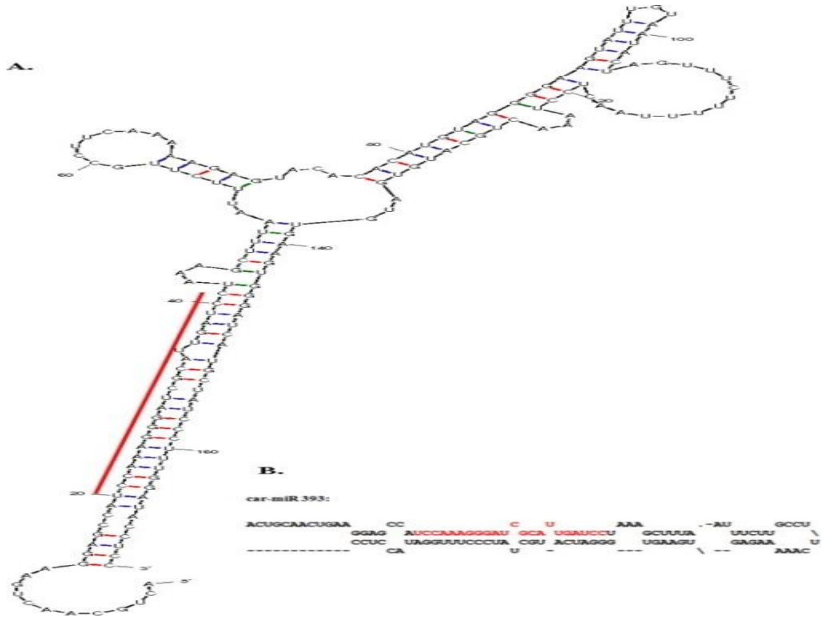

A.
B.
car-miR 393:
ACUGCAACUGAA CC C U AAA .-AU GCCU
GGAG AUCCAAAGGGAU GCA UGAUCCU GCUUUA UUCUU \
CCUC UAGGUUUCCCUA CGU ACUAGGG UGAAGU GAGAA U
------------ CA U - --- \ -- AAAC

Figure 3. Predicted potential miRNAs (source: Arzuba Akter et al.,2014)

The mfold web server is open to all users. No restrictions are applied to commercial users. However, users should be aware that the server is not secure and that data flowing both in and out may be detected by others. Moreover, query information is stored indefinitely in log files on the server. These log files are treated as confidential information, although gross statistics on usage are collected and disseminated. Furthermore, some of the submissions are selected as examples for teaching, but only if database searches reveal that the sequence is already in a public database (Michael Zuker, 2003).

3. POTENTIAL TARGET PREDICTION

After the identification of miRNAs, it is important to predict the potential targets which the identified miRNAs will act upon. There many tools available for the potential target prediction. Poonam et al. (2013) employed the miRanda program, which utilizes thermodynamics and dynamic programming alignments along with statistical parameters for target prediction in *Heliothis virescens* (Tobacco budworm). The parameters assigned for miRanda hybridization are default alignment score greater than or equal to 80, MFE of miRNA:mRNA duplex less than or equal to –37kcal/mol and the other parameters kept at default values. Dehury et al., (2013) used the Plant Small RNA Analysis Server (psRNATarget) http://plantgrn.noble. org/psRNATarget/ formerly known as miRU against the *Arabidopsis thaliana* DFCI Gene Index (AGI) Release 15.38 and ESTs of sweet potato to search for putative target genes to understand the biological functions of the newly identified sweet potato miRNAs. They used the following parameters: maximum exception of 0.5 (for lower false positive prediction), length of complementarity score: 20, target accessibility-allowed maximum energy to un-pair the target site (UPE): 25, flanking length around the target accessibility analysis: 17 bp upstream and 13 bp in downstream and range of central mismatch leading to translation inhibition: 9–11 nt. Further, the following criteria were set for identification of target genes: range of central mismatch for translation inhibition: 9–11 nt, a maximum exception value of 0.5, maximum mismatch at complementary site 3 without any gaps and the maximum target sites of 2. In addition to psRNA target server, the plant target prediction tool available at UEA srNA Tool Kit was also used for target prediction by following the guidelines of Schwab et al. (2006).

Dai and Zhao (2011) in a study applied psRNATarget program to search for the targets of identified miRNAs by homology algorithm. They also used *Arabidopsis* as a reference system for finding the targets of the candidate miRNAs in *Jatropha curcas*. They have the following criteria:1) Range of central mismatch for translational inhibition 9–11 nucleotide, 2) Maximum expectation value of 3, 3) Maximum mismatches at the complementary site ≤ 4 without any gaps, 4) Multiplicity of target sites 2 and the other parameters set with default: maximum expectation:2.0, length for complementarily scoring (hspsize): 20, target accessibility-allowed maximum energy to unpair the target site (UPE): 25.0, flanking length around target site for target accessibility analysis: 17 bp in upstream and 13 bp in downstream, and range of central mismatch leading to translation inhibition: 9–11nt.

4. CONCLUSIONS

All the aforementioned resources are very important for computational identification of miRNAs. However, these can be used for this purpose only if some initial data are present in publically available DNA repositories i.e. Genome sequences, GSS or EST sequences. Greater the number of reported EST or GSS of an organism higher will be the chances to predict new potential miRNAs by using bioinformatics methods. miRNA sequences predicted from the genomic sequences may not be authentic because sometimes these may not be expressed. Therefore majority of the researchers try to find the homologous sequences of the previously known miRNAs in the ESTs of the desired organism, cell or tissue because, occurrence of the candidate miRNA sequences in the ESTs confirms their expression and functions. Gaining some knowledge on the complete spectrum of miRNA identification using *in silico* approaches is crucial to understand the functions of miRNA and also to develop miRNA-based applications. Even though methods, as summarized in this review, have been developed in attempts to identify miRNAs, new algorithms are still in need to improve the ability to find new miRNAs and relate them with their respective functions.

5. REFERENCES:

Dehury Budheswar, Debashis Panda, Jagajjitsahu, Mousumisahu, Kishore sarma, Madhumita Barooah, Priyabratasen, and Mahendra Kumar Modi (2013). *In silico* identification and characterization of conserved miRNAs and their target genes in sweet potato (*Ipomoea batatas* L.) Expressed Sequence Tags (EST). Plant signaling & behavior 8:12

Gao Dadi, Robert Middleton,John E. J. Rasko and William Ritchie (2013). miREval 2.0: a web tool for simple microRNA prediction in genome sequences. Bioinformatics, 29(24). Pp 3225-3226.

Dai X., Zhao P.X. PsRNATarget: a plant small RNA target analysis server (2011). Nucleic Acids Res.;39:W155–W159. [PubMed].

Dyavegowda Padmashree and Narayanaswamy Ramachandra Swamy (2015). Computational identification of putative miRNAs and their target genes in pathogenic amoeba *Naegleria fowleri.* Bioinformation; 11(12): 550–557.

Lai, E.C., Tomancak, P., Williams, R.W., Rubin, G.M. (2003). Computational identification of Drosophila microRNA genes. Genome Biol. **4(7)**: R42.

Lee RC, Feinbaum RL, Ambros V. (2004). The *C. elegansheterochronic* gene line-4 encodes small RNAs with antisense complementarily to lin-14. Cell, 75(5):843-854.

Markham,N.R and Zuker, M. (2008). UNAfold: software for nucleic acid folding and hybridization. In Keith, J.M., editor, Bioinformatics, Volume II. Structure, function and applications, number 453 in Methods in Molecular Biology, Chapter 1, Pp.3-31.

Michael Zuker (2003). Mfold web server for nucleic acid folding and hybridization prediction., Nucleic Acids Research, 31(13): 3406–3415.

Poonam Chilana, Anu Sharma*, Vasu Arora, Jyotika Bhati & Anil Rai (2013). Computational identification and characterization of putative miRNAs in Heliothis virescens Bioinformation 9(2): 079-083.

Sam Griffiths-Jones,Harpreet Kaur Saini, Stijn van Dongen and Anton, J. Enright (2007). miRBase tools for microRNA genomic. Nucleic Acid Research, 36(1). Pp. 154-158.

Schwab R, Ossowski S, Riester M, Warthmann N, Weigel D.Plant Cell (2006). Highly specific gene silencing by artificial microRNAs in Arabidopsis.18(5):1121-33.

Wang Z. and Pesacreta TC. (2004). A subclassofmyosin XI associated with mitochondria, plastids, and the molecular chaperone subunits TCP 1alpha in maize. Cell Motil. Cytoskeleton 57, 218-232.

W. James Kent (2002). BLAT—The BLAST-Like Alignment Tool. Genomic Research, http://www.genome.org/cgi/doi/10.1101/gr.229202.

Wightman B and Ruvkun G. (1993). Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in *C. elegains*. Cell. 3;75(5). Pp 855-862.

William Ritchie, Francois-Xavier and Daniel Gautheret (2008). Mireval: a web tool for simple microRNA prediction in genome sequences. 24. Pp. 3094-3096.

Zhang J, (2006). Characterization of the transport mechanism and permanent binding profile of the uridine permease Fui1p of *Saccharomyces cerevisiae*. *J BiolChem* 281(38), 28210-21.

Zuker M (2003). Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res.1;31(13):3406-15.