Review Article

Bioinformatic Hunting of MicroRNAs

Iftekhar Ahmed Baloch¹ and Muhammad Din^{*1}

¹Department of Botany, University of Balochistan, Saria Road Quetta, Pakistan

^{*}Corresponding author email deen.2006@yahoo.com

Phone # +92-0333-7894859 Fax +92-081-9211-277

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. Presently there are three approaches for miRNA identification. Two approaches i.e. bioinformatic and small RNA library construction and sequencing are examples of reverse genetics approach. In these two approaches researchers try to find miRNAs in the known genotypes (sequences). A third approach called the classical approach is 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 bioinformatic 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. In 2001 the 1st release of miRBases (an online repository of miRNAs) contained only 218 miRNA entries and now its latest release 20 (released in June 2013) contains 24521 entries. The bioinformatics hunting of miRNAs has also a reasonable contribution in identification of new miRNAs. Around 25% of entries in miRBase (across all species) are predicted computationally. Likewise in Plant MicroRNA Database (PMRD) within the total 8433 miRNAs, 1297 of them are experimentally demonstrated and 7136 of them are from computational predictions. This review is aimed at describing the bioinformatic resources used in miRNA identification. Furthermore, the limitations of these resources are also described.

Keywords: Bioinformatics, Comparative genomics, MicroRNA

Abbreviations

Arabidopsis thaliana (ath)

Basic Local Alignment Search Tool (BLAST)

Complementary DNA (cDNA)

Double-stranded RNA (dsRNA)

Expressed sequence tags (ESTs)

Messenger RNAs (mRNAs)

MicroRNAs (miRNAs)

Multiple Sequence Alignment (MSA)

National Center for Biotechnology Information (NCBI)

Nucleotide (nt)

Plant Micro RNA Database (PMRD)

Populus trichocarpa (ptc)

Precursor miRNAs (pre-miRNA)

RNA induced silencing complex (RISC)

Transfer RNA (tRNA)

Untranslated regions (UTRs)

1. Introduction

MicroRNAs (miRNAs) are endogenous, small, noncoding RNAs of 18-25 nucleotide (nt) in length that negatively regulate their complementary messenger-RNAs (mRNAs) at the post transcriptional level in eukaryotic organisms [1]. By affecting the gene regulation, miRNAs are likely to be concerned with most biological processes [2]. Different sets of expressed miRNAs are found in different cell types and tissues [3]. In humans the non-coding RNAs such as miRNA, transfer RNA (tRNA) and others constitute 3% out of the total 5% functional genome [4]. Most miRNA genes are

found in intergenic regions or in anti-sense orientation to genes and contain their own miRNA gene promoter and regulatory units [5]. The genes of miRNAs are usually transcribed by RNA polymerase II [6]. The product is called a primary miRNA (Pri-miRNA) which may be thousands of nt in length and contains one or more miRNA stem loops [5]. The resulting transcript is capped with a specially modified nt at the 5' end, polyadenylated with multiple adenosines at 3' end [6] and spliced. The polymerase often binds to the promoter found near the DNA sequence encoding the hairpin loop of pre-miRNA. When a stem loop precursor is found in the

3' UTR, a transcript may serve as a pri-miRNA and an mRNA [7]. RNA polymerase III also transcribes some miRNAs especially those with upstream Alu sequences, tRNAs and mammalian wide interspersed repeat (MWIR) promoter units [8]. Primary transcripts of mature miRNAs fold into a stable stem loop structure forming the precursor miRNA (pre-miRNA). The pri-miRNAs are processed in the nucleus by the microprocessor complex, consisting of the RNase III enzyme Drosha, and the double-stranded RNA binding protein, Pasha/DGCR8 [9]. After being processed the pri-miRNA is changed into pre-miRNA. The resulting pre-miRNAs are approximately 70 nt in length in case of animals and in case of plants its length may be 90-140 nt. These are then folded into imperfect stem-loop structures. The pre-miRNA hairpins are exported to the cytoplasm where they are further processed into unstable, 18-25 nt miRNA duplex structures by the RNase III protein Dicer [10]. The less stable of the two strands in the duplex is incorporated into a multiple-protein nuclease complex, the RISC, which negatively regulates gene expression either by inhibiting translation elongation or by triggering mRNA destruction on the basis of the degree of complementarity of miRNA within its target [11,12].

2. MicroRNA identification approaches.

Presently there are three approaches for miRNA identification. Two approaches i.e. bioinformatics and small RNA library construction and sequencing are examples of reverse genetics approaches. In these two approaches researchers try to find miRNAs in the known genotypes (sequences). A third approach called the classical approach is example of forward genetics in which researchers determine the unknown genotype (miRNA sequence) of a known phenotype [13].

The majority of miRNAs are evolutionarily conserved from species to species within the same kingdom. The miRNA genes in one species may exist as orthologs in other species [14]. These are also conserved in animals and plants which suggests their early development in evolution [14]. Their conserved nature becomes an important logical tool for bioinformatics discovery of miRNAs in other species [15]. Since this review is aimed at describing the bioinformatic tools which are used in miRNA identification, the subsequent text will describe various bioinformatic resources and their limitations.

3. Bioinformatic resources and miRNA Research

Identification of miRNAs by using bioinformatic tools has now become one of the most widely used methods and it has facilitated the prediction of new miRNAs in both plant and animal systems [13]. This is largely used because of its low cost and high

efficiency. The logic behind this approach is finding homologous sequences of known miRNAs both within a single genome and across the genomes of related organisms [5]. There are some limitations of this homology search e.g. the major challenge is finding the miRNAs which are species specific and unrelated to the previously known organisms [13]. Sequence and structure homologies are used for computer based predictions of miRNAs. Computational strategies provide a valuable and efficient manner to predict miRNAs genes and their targets. The software based approach is applied to animals, human, fungi and plants [16]. By using bioinformatics tools Zhang et al., [17] identified 338 new possible miRNAs in 60 different plant species. Adai et al. [18] have predicted 43 new miRNAs in *Arabidopsis*. Li et al. [19] predicted 208 human pre-miRNA candidates from ESTs and introns Qiu et al. [20] reported miRNAs in cotton from ESTs. Barozai et al. [15] identified 22 miRNAs in cotton belonging to 13 miRNA families. Barozai et al. [21,22] identified 61 miRNAs in 7 species of *Helianthus* and 34 miRNAs in 2 species of *Mimulus* respectively by using ESTs.

The basic and important bioinformatics resources used in miRNA research are as follows.

3.1 MicroRNA Databases

With the discovery of a large number of miRNAs and their target genes, some public resources have been constructed e.g. miRBase [23]. (<http://microrna.sanger.ac.uk/>), provides a set of precursor and mature miRNAs discovered in many organisms. Similarly, *Arabidopsis thaliana* Small RNA Project (ASRP) (<http://asrp.cgrb.oregonstate.edu/>) lists miRNAs and their target genes in *Arabidopsis* [24]; Cereal small RNAs Database (CSRDB) (<http://sundarlab.ucdavis.edu/smrnas/>) is a collection of miRNAs identified in maize and rice [25]; Rfam (<http://rfam.sanger.ac.uk/>) provides secondary structures of miRNA precursors in many plant species [26]. The Plant Micro RNA Database (PMRD) (<http://bioinformatics.cau.edu.cn/PMRD/>) is a collection of 8433 miRNAs taken from 121 plant species [27].

Researchers from all over the world rely on these databases to check whether miRNA in a specific organism has been reported earlier or not. If miRNAs in a specific organism are not reported then consulting these databases creates an idea to report miRNA in that specific organism.

The important and main miRNA databases are miRBase and Plant Micro RNA Database (PMRD).

3.1.1 miRBase

miRBase database provides a searchable online storehouse for published miRNA sequences and associated annotation. miRBase also provides a gene naming and nomenclature function in the miRBase Registry [23]. The latest release 20 of the database (August, 2013) contains 24521 hairpin precursor miRNA sequences and 30424 mature miRNA sequences from 206 species, whereas Release 1.0 of the database (Dec, 2002) contained only 218 entries (Table. 1). This shows the enormous research and growth of data in the field of plant and animal miRNAs.

Both hairpin and mature miRNA sequences are available for searching and browsing, and entries can also be retrieved by name, keyword, references and annotation. All sequence and annotation data are also available for download. The data is publicly available through the web interface at <http://www.mirbase.org/> and in flat file form from <ftp://mirbase.org/pub/mirbase/>.

The miRBase database has following three main resources:

a. miRBase Registry provides a confidential service for the independent assignment of names to novel miRNA genes prior to their publication in peer-reviewed journals. Over 70 publications describing novel miRNA genes have made use of this service, and registration is a requirement of many journals. Researchers upload their miRNA sequences to this database before publication of novel miRNAs [13].

b. miRBase Sequences provides miRNA sequence data, annotation, references and links to other resources for all published miRNAs.

c. miRBase Targets provides an automated pipeline for the prediction of targets for all published animal miRNAs.

A limitation of miRBase is that it does not include the newly identified miRNA sequences by itself instead the researchers submit the sequences identified by them after their work gets published. Besides this the new versions are sometimes released very late leaving the researchers ignorant about the new sequences identified during the gap of two consecutive releases. For example there was a gap of almost a year between release 18 and 19 and same was the case of gap between release 19 and 20. However a minimum gap of one month was also observed in the previous releases.

3.1.2 Plant MicroRNA Database (PMRD)

Keeping in view the convenience of plant miRNA researchers, a plant microRNA database (PMRD) has been developed to combine the available data relevant to plant miRNAs from other public resources. The PMRD database web interface enables users to view and analyze each miRNA through either the Browse page or Search page. Detail

information page displays basic information for each unique miRNA. The evidence for each entry is either experimental or computational. In total, there are 8433 miRNAs taken from 121 plant species integrated into PMRD [27]. Within the total 8433 miRNAs in PMRD, 1297 of them are experimentally demonstrated and 7136 of them are from computational predictions. Users can view a miRNA display in the local Genome Browser in which users can extract information about *Arabidopsis*, rice and soybean. Through PMRD, users can also inspect miRNA expression profiles, target gene list and predicted interaction sites between miRNA and complementary DNA (cDNA). In addition, the secondary structure of miRNA precursor can be viewed by using the prediction tool of PMRD database. All available miRNA sequences in PMRD are downloadable.

In PMRD, there are 1427, 2540 and 2780 mature miRNA entries for *Arabidopsis*, rice and poplar, respectively, while in miRBase the entries for these three species are 207, 415 and 237, respectively. The other miRNAs from these three species mainly come from [28]. There are 166 soybean entries in PMRD, of which 79 sequences came from miRBase [23], and the others are from Zhang et al. [17,29]. For wheat there are deposited 85 miRNAs in PMRD, of which 32 came from miRBase [23] and the others are from Yao et al., [30], Zhang et al., [17] and Wei et al., [31]. As for maize, 207 entries are created in PMRD, including 98 from miRBase [23] and 109 from Zhang et al. [17]. For cotton there are 53 mature sequences in PMRD, in which 13 items were from miRBase [23], and the others are from Qiu et al., [20] and Zhang et al., [32].

3.2 miRNA homologue discovery tools

Since the bioinformatics approach is a reverse genetics based approach and usually its first step includes finding of homologue sequences of previously known miRNAs, there arises need of a bioinformatics tool which should fulfill this requirement. The mostly used bioinformatics tools for finding the orthologs sequences of previously known miRNAs are:

a. Basic Local Alignment Search Tool (BLAST)

In bioinformatics terminology BLAST stands for Basic Local Alignment Search Tool. The BLAST program was designed by Altschul et al., [33]. It enables the researchers to compare a query sequence with a library or database of sequences, and identify library sequences that resemble the query sequence above a certain threshold. For computational discovery of miRNAs, BLAST is performed to find the homologues of previously known miRNAs in the target organism. The query sequences (of previously

known miRNAs) are fetched from the public miRNA repositories such as miRBase and PMRD etc. Given a known miRNA (either a miRNA precursor sequence or mature miRNA sequence) as input, a researcher can find a set of candidate homologs. BLAST is also used for several other purposes. These include identifying species, locating domains, establishing phylogeny, DNA mapping, and comparison. One of the criteria for the validation of miRNAs is that they should not be protein coding and whether a nt sequence is protein coding or not, it can also be predicted by using Blastx [34].

b. Blast Like Alignment Tool (BLAT)

An extremely fast but considerably less sensitive alternative to BLAST that compares nt sequences to the genome is BLAT. Weber [14] found new human and mouse miRNA genes by homology search using BLAT.

c. BLASTZ

A version designed for comparing multiple large genomes or chromosomes is BLASTZ. Bentwich et al., [35] identified hundreds of conserved and non-conserved human miRNAs by using BLASTZ.

3.3 Multiple Sequence Alignment (MSA) Tools

In computational method of miRNA identification, the prediction of initial candidate pre-miRNAs is made by aligning a reference miRNA with its corresponding raw sequences, genomic sequences and the aligned portion of the raw sequence is considered as an initial candidate miRNA. For this purpose MSA tools such as Clustal W [36] and Clustal X [37] are used. By the late 1990s, Clustal W and Clustal X were the most widely used multiple alignment programs. They were able to align medium-sized data sets very quickly and were easy to use. The alignments were of sufficient quality not to require manual editing or adjustment very often. This situation changed greatly with the appearance of the first custom made standard test set for multiple alignment programs, Benchmark Alignment dataBASE (BALiBASE) [38]. This was followed by the appearance of T-Coffee which was able to make very accurate alignments of very divergent proteins but only for small sets of sequences, given its high computational cost. With the increase in processing speed of desktop computers, and subsequent optimisation of the T-Coffee code, the latter is now practical for routine use on moderately sized

alignment problems. More recently, MAFFT and Multiple Sequence Comparison by Log- Expectation (MUSCLE) appeared; which were, initially, at least as accurate as Clustal, in terms of alignment accuracy, but which were also extremely fast; and able to align many thousands of sequences. Over the past few years, these programs have also gradually become more and more accurate with difficult alignments. Nonetheless, Clustal W and Clustal X continue to be very widely used, increasingly on websites. The European Bioinformatics Institute (EBI) Clustal site gets literally millions of multiple alignment jobs per year.

3.4 Conservation analysis tools

The miRNAs are evolutionarily conserved from species to species within the same kingdom. The miRNA genes in one species may exist as orthologs or homologs in other species [14]. Researchers are interested to study the conserved miRNAs in different organisms. Multiple sequence alignment tools such as Clustal W (<http://www.ebi.ac.uk/Tools/msa/clustalw2>) [36] and Web logo (<http://weblogo.berkeley.edu/logo.cgi>) [39] can also be used for conservation analysis of miRNAs.

3.5 Secondary structure prediction tools.

A miRNA assumes a specific hairpin secondary structure before moving to cytoplasm. In order to predict the secondary structure of miRNAs several bioinformatics tools may be used e.g. PMRD [28] and micro PC can predict the secondary structure of a pre-miRNA sequence. The most commonly used RNA secondary structure prediction tools is mfold [40]. 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. it also predicts the thermodynamic details for folded structures e.g. minimum folding energy (mfe). Besides this mfold allows the users to redraw the predicted structure.

3.6 Target identification

The identification of miRNA targets is an important step for understanding their regulatory function. Different computational approaches have been developed and used to predict their targets [41]. Most of the computational tools developed to identify mRNA target sequences depend greatly on complementarity between the miRNA seed sequence (nt 2-7) and the target sequence [42]

Table.1. Number of miRNA (hairpin) entries in various releases of miRBase.

Release	Entries	Release Date	Gap from previous release	New microRNAs added
1.0	218	Dec.2002	-	-
1.1	262	Jan.2003	1 month	44
1.2	295	Apr.2003	3 months	33
1.3	332	May.2003	1 month	37
1.4	345	Jul.2003	2 months	13
2.0	506	Jul.2003	1 month	161
2.1	558	Sep.2003	2 months	52
2.2	593	Nov.2003	2 months	35
3.0	719	Jan.2004	2 months	126
3.1	899	Apr.2004	3 months	170
4.0	1185	Jul.2004	3 months	286
5.0	1345	Sep.2004	2 months	160
5.1	1420	Dec.2004	3 months	75
6.0	1650	Apr.2005	4 months	230
7.0	2909	Jun.2005	2 months	1259
7.1	3424	Oct.2005	4 months	515
8.0	3518	Feb.2006	4 months	94
8.1	3963	May.2006	3 months	445
8.2	4039	Jul.2006	2 months	76
9.0	4361	Oct.2006	3 months	322
9.1	4449	Feb.2007	4 months	88
9.2	4584	May.2007	3 months	135
10.0	5071	Aug.2007	3 months	487
10.1	5395	Dec.2007	4 months	324
11.0	6396	Apr.2008	4 months	1001
12.0	8619	Sep.2008	5 months	2223
13.0	9539	Mar.2009	6 months	920
14	10883	Sep.2009	6 months	1344
15	14197	Apr.2010	7 months	3314
16	15172	Aug.2010	4 months	975
17	16772	Apr.2011	8 months	1600
18	18226	Sep. 2011	5 months	1454
19	21264	Aug. 2012	11 months	3038
20	24521	Jun.2013	10 months	3257

Diana-microT [43] was one of the first computational tools for target prediction that identified specific interaction rules based on bioinformatics and experimental approaches. The tool successfully recovered all validated *C. elegans* miRNA targets.

Several additional methods for the prediction of miRNA targets have been subsequently developed. These methods mainly use sequence complementarities, thermodynamic stability calculations and evolutionary conservation among species to determine the likelihood of formation of a productive miRNA–mRNA duplex [1,44]. John et al. [45] developed the miRanda algorithm for miRNA target prediction. The miRanda uses dynamic programming to search for optimal sequence complementarities between a set of mature miRNAs and a given mRNA. miRNA.org (<http://www.microrna.org>) [46] is a comprehensive resource of miRNA target predictions and miRNA ‘expression profiles’. Target predictions are based on the miRanda algorithm, whereas miRNA ‘expression profiles’ are derived from a comprehensive sequencing project of a large set of mammalian tissues and cell lines of normal and disease origin.

Brennecke et al. [47] have suggested that the 3′ out-seed segment of the miRNA–mRNA duplex can compensate for imperfect base pairing of the target with the seed segment, and a recent computational approach [48] has considered the contributions of both seed and out-seed miRNA segments in target identification. Using sequence conservation reduces false-positive predictions but, as a result, some less-conserved target sites may be missed. This presents a dilemma, which is how to avoid rejection of these less highly-conserved target sites while still reducing the very large numbers of predictions that are found when seed region conservation in the target is not required. In order to reduce the false-positive predictions inherent in methods that heavily weight specific target sequence conservation, Lewis et al. [49] developed TargetScanS. TargetScanS scores target sites based on the conservation of the target sequences between five genomes (human, mouse, rat, dog and chicken) because evolutionarily conserved target sequences are more likely to be true targets. In testing, TargetScanS was able to recover targets for all 5300 human genes known at the time to be targeted by miRNAs.

PicTar [50] is a computational method used to detect common miRNA targets in vertebrates, *C. elegans* and *Drosophila*. PicTar is based on a statistical method applied to eight vertebrate genome-wide alignments (multiple alignments of orthologous nucleotide sequences (3′-UTRs). PicTar was able to recover validated miRNA targets at an estimated 30% false-positive rate. In a separate study, PicTar was

applied to target identification in *Drosophila melanogaster* [51]. These studies suggest that one miRNA can target 54 genes, on average, and that known miRNAs are projected to regulate a large fraction of all *Drosophila melanogaster* genes (15%). This is likely to be a conservative estimate because of the incomplete input data.

TargetBoost [52] is a machine-learning algorithm for miRNA target prediction using only sequence information to create weighted sequence motifs that capture the binding characteristics between miRNAs and their targets. Sung et al. [44] also reported the development of a machine-learning algorithm using SVM. The best reported results were 0.921 sensitivity and 0.833 specificity. [48] used a machine-learning approach that employs features extracted from both seed and outseed segments. The best result obtained was an accuracy of 82.95%, which was generated using only 48 positive and 16 negative human examples— a relatively small training set to assess the algorithm.

Thadani & Tammi (2006) chunked MicroTar, a new statistical computational tool for prediction of miRNA targets from RNA duplexes, which does not use sequence homology for prediction. MicroTar mainly relies on a quite novel approach to estimate the duplex energy. However, the reported sensitivity (60%) is significantly lower than that achieved using other published algorithms. At the same time, a miRNA pattern-discovery method, RNA22 [53], was proposed for use in scanning UTR sequences for targets. RNA22 does not rely upon cross-species conservation but was able to recover most of the known target sites with validation of some of its new predictions.

Yousef et al. [42] described a target-prediction method (NBmiRTar) using, instead machine learning by a Naïve Bayes classifier. NBmiRTar does not require sequence conservation but generates a model from sequence and miRNA–mRNA duplex information derived from validated target sequences and artificially generated negative examples. In this case, both the seed and the ‘out-seed’ segments of the miRNA–mRNA duplex are used for target identification. The NBmiRTar technique produces fewer false-positive predictions and fewer target candidates to be tested than miRanda [45]. It exhibits higher sensitivity and specificity than algorithms that rely only on conserved genomic regions to decrease false-positive predictions.

In 2004, in a review, Lai [41] noted that there was almost no overlap among the predicted targets identified by the various methods and suggested that each tool captures a subset of the entire target class as a function of the specific features they have incorporated into their prediction models. Sethupathy

et al. [54] conducted a comparison of the five most commonly used tools for mammalian target prediction. This study indicated that 30% of the experimentally validated target sites are nonconserved, supporting the need for the development of different or complementary computational approaches to capture new target sites. Furthermore, the large number of predictions that each of these tools is producing suggests that the heavy reliance on homology or comparative-sequence analysis is not sufficient to generate accurate predictions with a high sensitivity and there are yet-to-be identified recognition parameters that must be considered.

Many of the bioinformatics tools were designed for animal miRNAs which are considerably different from plant miRNAs in the target binding process. Some of these differences are loose complementarity hybridization, inhibition of target gene's expression at the translational level and hybridization with the 3'-UTR region of its target mRNA. Whereas, plant miRNAs are perfectly or near perfectly hybridized with mRNAs, degradation of target mRNA and can recognize and hybridize to any region of the mRNA. Dai and Zhao (2011) [55] presented a new plant small RNA target analysis web server, psRNATarget (<http://plantgrn.noble.org/psRNATarget/>) and replace the miRU. The psRNATarget is now widely used for miRNA target prediction in plants.

Another algorithm, RNA-Hybrid [56] is similar to an RNA secondary-structure prediction algorithm, like the MFOLD program [40] but it determines the most favorable hybridization site between two sequences. The algorithmic core of RNAhybrid is a variation of the classic RNA secondary structure prediction [57]. Instead of a single sequence that is folded back onto itself in the energetically most favourable fashion, RNA-Hybrid determines the most favourable hybridization site between two sequences. Though in principle these two sequences can be arbitrarily long, for miRNA target prediction, the target candidate will be rather long (hundreds to thousands of nt) and the miRNA will be between 19 and 24 nt. Since miRNA/target interactions have not been reported to contain bifurcations (also called multi-loops), these are not considered by RNAhybrid, thus considerably increasing the speed of the algorithm. RNA-Hybrid does not use any RNA folding or pairwise sequence alignment code, but implements an algorithm that was specifically designed for RNA hybridization.

Recently an interesting approach for miRNA target prediction is introduced by Barozai [58]. In this approach, the miRNAs are submitted as queries to the NCBI Basic local alignment search tool program with the following parameters: (1) database reference: RefSeq RNA; (2) organism: interested organism; and

(3) program selection: highly similar sequences (megablast). Afterwards, the mRNA sequences showing 75% query coverage are selected and submitted to RNAhybrid for the confirmation of the potential targets. Only targets having stringent seed site located at positions 2–7 from the 5' end of the miRNA along or with the supplementary site are selected and further submitted to the another prediction tool like; miRanda to reconfirm the RNAhybrid results.

4. Conclusion

All the above mentioned resources are very important but these can be used for the bioinformatic identification of miRNAs only if some initial data is present at Publically available DNA repositories i.e. Genome sequences, GSS or EST sequences. Greater the number of reported EST or GSS sequences of an organism higher will be the chances to predict new potential miRNAs by using bioinformatic methods. miRNA researchers unearth the meaningful information for the miRNAs from this raw data (GSS and EST sequences). miRNA sequences predicted from the genomic sequences may not be authentic because sometimes these are not 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.

References

1. Bartel DP (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* (116):281–297.
2. Lim LP, Lau NC, Garrett-Engle P, Grimson A, Schelter JM et al., (2005). Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature* 433: 769-773.
3. Lagos-Quintana M, Rauhut R, Yalcin A, Meyer J, Lendeckel W et al., (2002). Identification of tissue-specific microRNAs from mouse. *Curr Biol.* 12 (9): 735–739.
4. Human Genome Sequencing Consortium. Initial sequencing and analysis of human genome, *Nature* 420 (2002) 520-562.
5. Lee RC, Ambros V (2001) An extensive class of small RNAs in *Caenorhabditis elegans*, *Science*, vol. 294, no. 5543, 862–864.
6. Lee Y, Kim M, Han J, Yeom KH, Lee S et al., (2004) MicroRNA genes are transcribed by RNA polymerase II. *EMBO J.* 23. 4051–4060.
7. Cai X, Hagedorn CH, Cullen BR (2004) Human microRNAs are processed from capped, polyadenylated transcripts that can also function as mRNAs. *RNA* 10: 1957–1966.

8. Faller M, Guo F (2008) MicroRNA biogenesis: There's more than one way to skin a cat. *Biochimica et Biophysica Acta* 1779 (11):663–667.
9. Denli AM, Tops BB, Plasterk HA, Ketting RF, Hannon GJ (2004) Processing of primary microRNAs by the Microprocessor complex. *Nature*. 432(7014):231–235.
10. Sontheimer EJ (2005) Assembly and Function of RNA Silencing Complexes. *Nature Reviews. Molecular Cellular Biology*. 6: 127–138.
11. Aukerman MJ, Sakai H (2003) Regulation of flowering time and floral organ identity by a microRNA and its APETALA2-like target genes. *Plant Cell* 15:2730–2741.
12. Tang G, Reinhart BJ, Bartel DP, Zamore PD (2003) A biochemical framework for RNA silencing in plants. *Genes Dev* 17:49–63.
13. Unver T, Namuth-Covert DM, Budak H (2009) Review of current methodological approaches for characterizing MicroRNAs in plants. *International Journal of Plant Genomics*, vol. 2009, Article ID 262463, 11 pages.
14. Weber MJ (2005) New human and mouse microRNA genes found by homology search. *FEBS J* 272: 59–73.
15. Barozai MYK, Irfan M, Yousaf R, Ali I, Qaisar U et al., (2008) Identification of micro-RNAs in Cotton. *Plant Physiology and Biochemistry* 46(8–9):739–751.
16. Lewis BP, Shih IH, Jones-Rhoades MW, Bartel DP, Burge CB (2003) Prediction of mammalian MicroRNA targets. *Cell* 115 (7) 787–798.
17. Zhang BH, Pan XP, Wang QL, Cobb GP, Anderson TA (2005) Identification and characterization of new plant microRNAs using EST analysis. *Cell Res* 15:336–360.
18. Adai A, Johnson C, Mlotshwa S, Archer-Evans S, Manocha V et al., (2005) Computational prediction of miRNAs in *Arabidopsis thaliana*. *Genome Research* 15 (1):78–91.
19. Li, S. C., Pan, C. Y., & Lin, W. C. (2006). Bioinformatic discovery of microRNA precursors from human ESTs and introns. *BMC genomics*, 7(1), 164.
20. Qiu CX, Xie FL, Zhu YY, Guo K, Huang SQ et al., (2007) Computational identification of microRNAs and their targets in *Gossypium hirsutum* expressed sequence tags. *Gene* 395:49–61.
21. Barozai, MYK, Baloch IA, Din M (2011a). Identification of MicroRNAs and their targets in *Helianthus*. *Mol Biol Rep*. 39: 2523–2532.
22. Barozai MYK, Din M, Baloch IA (2011b). Identification of microRNAs in ecological model plant *Mimulus*. *Journal of Biophysical Chemistry*, 2(3): 322–331.
23. Griffiths-Jones S (2004) The microRNA Registry. *Nucleic Acids Res* 32D:109–111.
24. Gustafson AM, Allen E, Givan S, Smith D, Carrington JC et al., (2005) ASRP: The *Arabidopsis* Small RNA Project Database. *Nucleic Acids Res*. D637–640.
25. Johnson C, Bowman L, Adai AT, Vance V, Sundaresan V. (2007). CSRDB: a small RNA integrated database and browser resource for cereals. *Nucleic Acids Res*, 35:D829–33.
26. Gardner PP, Daub J, Tate JG, Nawrocki EP, Kolbe DL et al., (2009) Rfam: updates to the RNA families database.
27. Zhang Z, Yu J, Li D, Zhang Z, Liu F et al., (2010) PMRD: plant microRNA database. *Nucleic Acids Res*. D806–13.
28. Lindow M, Jacobsen A, Nygaard S, Mang Y, Krogh A et al., (2007) Intragenomic matching reveals a huge potential for miRNA-mediated regulation in plants. *PLoS Comput. Biol.* 3:e238.
29. Zhang B, Pan X, Stellwag EJ (2008) Identification of soybean microRNAs and their targets. *Planta* 229 :161–182.
30. Yao Y, Guo G, Ni Z, Sunkar R, Du J et al., (2007) Cloning and characterization of microRNAs from wheat (*Triticum aestivum* L.) *Genome Biol* 8:R96.
31. Wei B, Cai T, Zhang R, Li A, Huo N, Li S, Gu YQ, Vogel J, Jia J, Qi Y, et al., (2009) Novel microRNAs uncovered by deep sequencing of small RNA transcriptomes in bread wheat (*Triticum aestivum* L.) and *Brachypodium distachyon* (L.) *Beauv. Funct. Integr. Genomics.*; 9:532–541.
32. Zhang B, Wang Q, Wang K, Pan X, Liu F et al., (2007) Identification of cotton microRNAs and their targets. *Gene*, vol. 397, no. 1–2, 26–37.
33. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215:403–410.
34. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zheng Z et al., (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res*. 25:3389–3402.
35. Bentwich I, Avniel A, Karov Y, Aharonov R, Gilad S, Barad O and Bentwich Z (2005) Identification of hundreds of conserved and nonconserved human microRNAs. *Nature genetics*, 37 (7): 766–770.
36. Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic acids research*, 22 (22) :4673–4680.
37. Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality

analysis tools. *Nucleic acids research*, 25 (24) :4876-4882.

38. Thompson JD, Anne Bahr, Thierry JC, Poch O (1999) BALiBASE: a benchmark alignment database for the evaluation of multiple alignment programs. *Bioinformatics* 15 :87–88.

39. Crooks GE, Hon G, Chandonia JM, Steven E. Brenner (2004) Web- Logo: a sequence logo generator. *Genome Res* 14:1188–1190.

40. Zuker M (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res* 31:3406–3415.

41. Lai, E. C. (2004). Predicting and validating microRNA targets. *Genome biology*, 5, 2004-2005.

42. Yousef M, Jung S, Kossenkova AV, Showe LC, Showe MK (2007) Naive Bayes for microRNA target predictions machine learning for microRNA targets. Oxford University Press, 2987-2992.

43. Kiriakidou M, Nelson PT, Kouranov A, Fitziev P, Bouyioukos C et al., (2004) A combined computational-experimental approach predicts human microRNA targets. *Genes & Development* 18: 1165-1178.

44. Sung KK, Jin WN, Wha JL, Byoung TZ (2005) A Kernel method for microRNA target prediction using sensible data and position-based features. In Proceedings of the 2005 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology. 46–52.

45. John B, Enright AJ, Aravin A, Tuschl T, Sander C et al., (2004) Human MicroRNA Targets. *PLoS Biology*; 2, e363.

46. Betel D, Wilson M, Gabow A, Marks DS, Sander C (2008). The microRNA.org resource: targets and expression. *Nucleic acids research*, 36 (suppl 1): D149-D153.

47. Brennecke J, Stark A, Russell RB, Cohen SM (2005) Principles of microRNA-target recognition. *PLoS Biol* 3, e85.

[48] Yan X, Chao T, Tu K, Zhang Y, Xie L et al., (2007) Improving the prediction of human

microRNA target genes by using ensemble algorithm. *FEBS Letters*; 581: 1587.

49. Lewis B, Burge C, Bartel D (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* 120(1) :15-20.

50. Krek A, Grün D, Poy MN, Wolf R, Rosenberg L et al., (2005) Combinatorial microRNA target predictions. *Nat Genet* 37: 495-500.

51. Grün D, Wang YL, Langenberger D, Gunsalus KC, Rajewsky N (2005) microRNA target predictions across seven *Drosophila* species and comparison to mammalian targets. *PLoS Comput Biol* 1, e13.

52. SaeTrom OLA, Snove OJ , SaeTrom PAL (2005) Weighted sequence motifs as an improved seeding step in microRNA target prediction algorithms. *RNA* 11 :995–1003.

53. Miranda KC, Huynh T, Tay Y, Ang YS, Tam WL et al., (2006) A patternbased method for the identification of microRNA binding sites and their corresponding heteroduplexes. *Cell* 126:1203–1217.

54. Sethupathy P, Megraw M, Hatzigeorgiou AG (2006) A guide through present computational approaches for the identification of mammalian microRNA targets. *Nat. Methods* (11) 881-886.

55. Dai, X., & Zhao, P. X. (2011). psRNATarget: a plant small RNA target analysis server. *Nucleic acids research*, 39(suppl 2), W155-W159.

56. Rehmsmeier M , Steffen P, Höchsmann M and Robert G (2004). Fast and effective prediction of microRNA/target duplexes. *RNA*, 10:1507-1517.

57. Zuker M, Stiegler P (1981) Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.* 9:133–148.

58. Barozai, M. Y. K. (2012). Identification and characterization of the microRNAs and their targets in *Salmo salar*. *Gene*, 499 (1), 163-168.