International Journal of Applied Research on Information Technology and Computing
DOI: 10.5958/0975-8089.2017.00003.3 Indianjournals.com

Computational Algorithm for Prediction of miRNA in Plants

H. Ravisankar[1]*, K. Sivaraju[2], K. Prabhakar Rao[3], K. Sarala[4] and D. Damodar Reddy[5]

[1]Principal Scientist, Department of Computer Applications, [2]Principal Scientist, Department of Division of Crop Chemistry and Soil Science; [3]Scientist, [4]Head, Department of Division of Crop Improvement, [5]Director, Department of ICAR-Central Tobacco Research Institute, Rajahmundry-533105, AP, India
**Corresponding author email id: hravisankar@india.com*

ABSTRACT

The role of bio-informatics and computational biology in life sciences has been growing ever since the emergence of complex and large datasets for understanding the biological processes and expression of traits. Even though, algorithmic approaches are available, individual modules have to be executed for each intermediate result to predict the microRibosomal Nucleic Acid (miRNA). Hence, an attempt was made to develop an integrated model for predicting the miRNA in which all the structures will be generated automatically, once we submit the genomic sequences with varied datasets as an input. A novel algorithm was developed for prediction of miRNA in plants using shell scripting for fast processing of huge amount of data. As a part of the pipeline, software modules for generating RNA secondary structure, RNA structure in Extensible Markup Language(XML) format and RNA structure in pictorial view were developed using shell scripting by imposing various constraints, namely(1) miRNA should be a part of hairpin, (2) miRNA length is approximately 21nt,(3) it should start from 41st position and (4) the length of hairpin of good miRNA is >50 nt. Built-in modules, namely 'samtools' and 'mfold' were used in the scripting for generating RNA secondary structure in graphical form and in XML format. These modules were executed with the representative tobacco genome survey sequence and able to retrieve the above structures which are considered an input for predicting miRNA, and an output file was generated to display good miRNA sequences from the given structure. This algorithm can be used for predicting miRNA from the genomic sequences from the upcoming tobacco and other plantgenome projects.

Keywords: Algorithm, Dataset, miRNA, Module, Sequence, Software, Secondary structure

1. INTRODUCTION

MicroRNAs (miRNAs) are a new class of non-coding endogenous RNA molecules,which play crucial roles as regulators of gene expressions in eukaryotes. The first endogenous ~22nt RNAs identified were lin-4 RNA and let-7 RNA, both of which involved in controlling the timing of larval development in the nematode *Caenorhabditiselegans*[1,2]. Processed from primary transcripts that are transcribed from miRNA genes, mature miRNAs are usually 19–25 nucleotides long. Mature miRNAs are thought to down regulate the translation of messenger RNAs after recognising and binding to partially complementary sites in the 3′ untranslated regions of the messenger RNA.

miRNAs usually participate in a set of important life process, including growth processes, haematopoiesis, organ formation, apoptosis and cell proliferation. They are also closely related to many kinds of human diseases, including cancer[3]. Therefore, their study is important for the understanding of various cell functions in eukaryotic species. miRNAs have been discovered by various experimental methods, such as, northern blot[4,5], clone library[6] and separation of microRNAs[7]. These methods are highly biased towards abundantly expressed miRNAs and only abundant miRNA genes can be detected[8]. It is imperative that not all miRNAs

are well expressed in many tissues, cell types and development stages that have been tested[9]. However, the miRNA identification by experiments is time consuming and cost expensive. The algorithms used for gene prediction are less efficient to predict miRNAs because of the low similarity existed among the miRNA sequences. A number of computational algorithms were developed to predict miRNA with respect to those of other species miRNA precursors[10]. The miRNAs can be detected using characteristics such as the secondary structure and free folding energy of their precursors, conservation a part of miRNA sequences or similarity with other miRNAs. These characters have been exploited widely in the development of miRNA finding algorithms[11,12]. Precursor miRNA (pre-miRNA) of 60–70nt have stem-loop hairpin structures, which are an important characteristic feature used in the computational identification of miRNAs. Recently, the ab initio method based on machine learning was developed and applied to distinguish real pre-miRNAs from candidate hairpin sequences. Through learning from known miRNAs and pre-miRNAs, the features of primary sequence and secondary structure are extracted. These features are used to construct classifiers, such as support vector machine (SVM)[12]. MiPred[13] is the extension of Triplet-SVM, which uses two additional features such as minimum of free energy and the randomisation test (*P* value), totalling 34 features. Xuan *et al.*[14] proposed a novel feature selection method based on genetic algorithm, according to the characteristics of human pre-miRNAs, which improved the accuracy nearly 12% compared with MiPred.

All these methods relied mainly on evolutionary conservation to eliminate a large number of false-positive predictions. However, a substantial number of lineage- or species-specific miRNA genes do exist which escape the prediction of conservation-based approach[15]. If the miRNA precursors of one species have been not known, the methods are impossible to predict putative miRNA precursors in other similar species. The importance of miRNAs in the post-transcriptional regulation, the lack of sufficient number of known miRNAs and poorly annotated genomes collectively necessitated for the novel effective computational approaches for miRNA prediction. In this work, we developed a novel algorithm to identify miRNA from plant genome sequences.

In this paper, we developed a novel algorithm for prediction of miRNA in plant genomeusing shell scripting for fast processing of huge amount of data.

2. BACKGROUND

The discovery of miRNAs represents one of the most significant advances in biological and medical sciences in the last decade. Hundreds of miRNAs have been identified in plants, viruses, animals and humans, and these tiny, non-coding RNA transcripts have been found to play crucial roles in important biological processes involved in human health and disease. The term miRNA was first introduced in a set of three articles in Science (26 October 2001). Although the first published description of miRNA appeared in 1993[1], only in the last few years has the breadth and diversity of this class of small regulatory RNAs been appreciated. A great deal of effort has gone into understanding how, when and where miRNAs are produced and their functions in cells, tissues and organisms. Each miRNA is thought to regulate multiple genes. Ashundreds of miRNA genes are predicted to be present in higher eukaryotes, the potential regulatory circuitry afforded by miRNA is enormous. Several research groups have provided evidence that miRNAs may act as key regulators of processes as diverse in early development[2], cell proliferation and cell death[16], apoptosis and fat metabolism[17] and cell differentiation[18]. Recent studies of miRNA expression implicated that miRNAs play an important role in brain development[19], chronic lymphocytic leukaemia[20], colonic adenocarcinoma, Burkitt's lymphoma and viral infections[21] suggesting possible links between miRNAs and viral disease, neurodevelopment and cancer. There is speculation that in higher eukaryotes, the role of miRNAs in regulating gene expression could be as important as that of transcription factors.

3. MATERIALS AND METHODS

The algorithm was developed for prediction of new miRNA from plant genome based on the conditions, namely (1) miRNA should be a part of hairpin, (2) miRNA length is approximately 21nt, (3) it should start from 41 stposition and (4) the length of hairpin of good miRNA is >50nt. The software was developed using shell scripting under Linux environment.

For analysing the homologous miRNAs in plants, downloaded Genome Survey Sequences (GSS) sequences of tobacco and analysis were conducted with 100 representative sequences. The length of the hairpin motif varies and also a miRNA can originate from the 5p or 3p end of the hairpin motif. To account for the variation in the length and position of the mature miRNA within the hairpin sequence, we used flanking sequences of different lengths for each match to the mature sequence. From the starting position of the mature match to the genome sequence, the following coordinates are used to obtain the flanking sequences (-10, +60; -20, +80;-20, +120;-40, +90). The secondary structure folding for these sequences is obtained using 'mfold'. Each of the resulting candidate secondary structures with the least free energy are evaluated for

(a) the mature sequence resides on either one of the 5p or 3p arms,

(b) the mature sequence does not continue into the hairpin loop and

(c) the structure does not contain additional internal bulge loops that form another secondary structure within the hairpin structure.

From the candidates, thus obtained from running the four different flanking sequences, the sequence with the least free energy is presented. The mature sequence can originate either from both 5p and 3p ends, and for each candidate both feasible options were obtained.

The sequence of steps implemented is as follows:

1. Generate the sequence for each chromosome of data file.
2. Generate the RNA structure for each sequence. The output for each sequence creates a folder containing various files. Three files which are important among them are (1) a file with '.det' extension is created which contains the RNA secondary structure with details that includes the fine features such as helix loop, hairpin loop, multi loop, interior loop and others. (2) A file with extension '.rnma' is created which specifies the structure of RNA in XML format and (3) the file '.out' is created which contains the structure of RNA in graphical view.
3. Two files '.det' and '.rnma' are compared with various conditions like (1) miRNA should be a part of hairpin, (2) miRNA length is approximately 21nt, (3) it should start from 41st position and(4) the length of hairpin of good miRNA is >50nt.

4. RESULTS AND DISCUSSION

To analyse the new datasets for prediction of miRNA, the following sequence of steps isto be followed. The code was developed using Java, Shell Script and Perl in Linux environment. The results were generated by taking one input example 'Chr5-26236044'.

In step 1, the shell programme takes the input as given chromosome that is Chr5-26236044 and generates a range by subtracting -20 as left coordinate and adds +120 as right coordinate.

4.1 Step 1

sh ranges.sh

(This programme defines a range for each chromosome. Left coord: -20; Right coord: +120)

Output example:Chr5:26236024-26236184

Output filename: dataset1

The output in step 1 that is 'dataset1' which consists of 'Chr5:26236024-26236184' as input for step2 and this shell generates a genome sequence named 'mirnainput1'.

4.2 Step 2

sh seq1

(This programme generates a sequence for each chromosome)

Input filename: dataset1

Output filename: mirnainput1

Output example:

>Chr5:26236024-26236184

CGCCGTCGCCACCGCCGCCGCCTGCCGCGTAGTCGTACTTGAAACCGAGCGCTGGCGGCC

CCGACGGCTCCAGCGGCAGCAGCGCTGCCCCGGGCCCGACTCCTGGGCCGGGATCGCGCC

CGCGCTCCTCACGCTTCAGGCCCCCGCCGCCCTCGGCGCAC

The generated genome sequence is considered an input for step3. In this step, different structures for the given sequence will be generated where the miRNA was found, namely structure 1, structure 2, structure 6, structure 7, structure 8, structure 9, structure 10 and structure 12.

4.3 Step 3

sh main.sh

(This shell includes two programmes: final8 and RNA.java)

(This programme predicts the good miRNAs. Input filename: mirnainput1/(give the filename in main.sh)

In 'final8' programme, give the range as '20'output filename: MIRNA1.out

Chr5:26236024-26236184

-72.30

41:62

-71.90

41:62

120:141

-71.90

-71.30

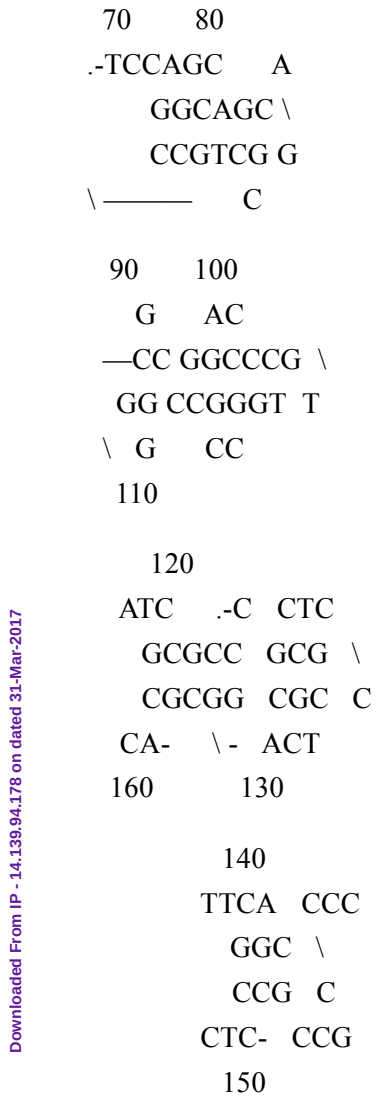
-71.10

-71.00

41:62
70:91
105:126
72:93
-70.60
41:62
92:113
-69.60
20:41
47:68
82:103
123:144
92:113
-69.60
41:62
-69.00:
41:62
70:91
110:131
72:93
78:99
81:102
-69.00:
-68.90:
41:62
73:94
-68.90:
Structure 1

Folding bases 1 to 161 of Chr5:26236024-26236184
Initial dG = -72.3

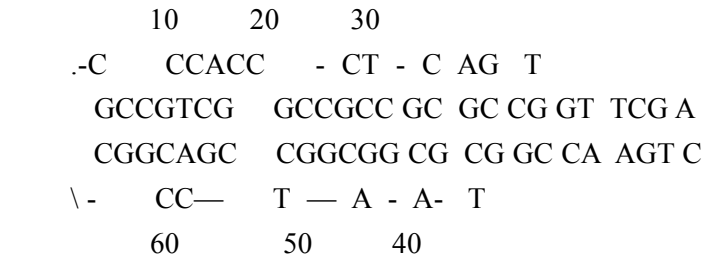
```
         10         20         30
.-C|      CCACC      -  CT  -  C  AG   T
  GCCGTCG     GCCGCC GC  GC CG GT  TCG A
  CGGCAGC     CGGCGG CG  CG GC CA  AGT C
\ -^       CC—       T  —  A  -  A-   T
         60          50         40
```

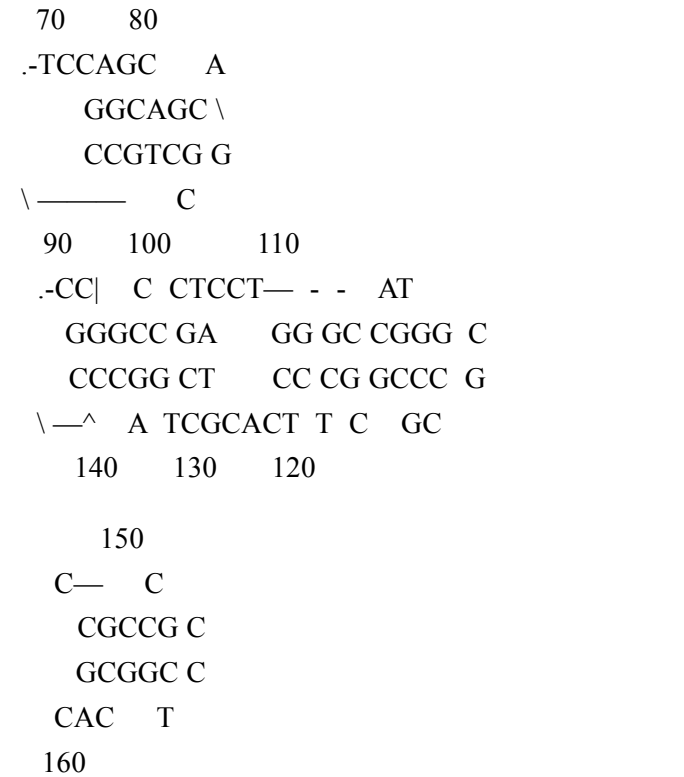


In this step, 13 secondary RNA structures were generated. Out of them, secondary structure of miRNA was found in the structure 8 only. Hence, structures 1 and 8 were given.

Structure 2

Folding bases 1 to 161 of Chr5:26236024-26236184
Initial dG =-71.9

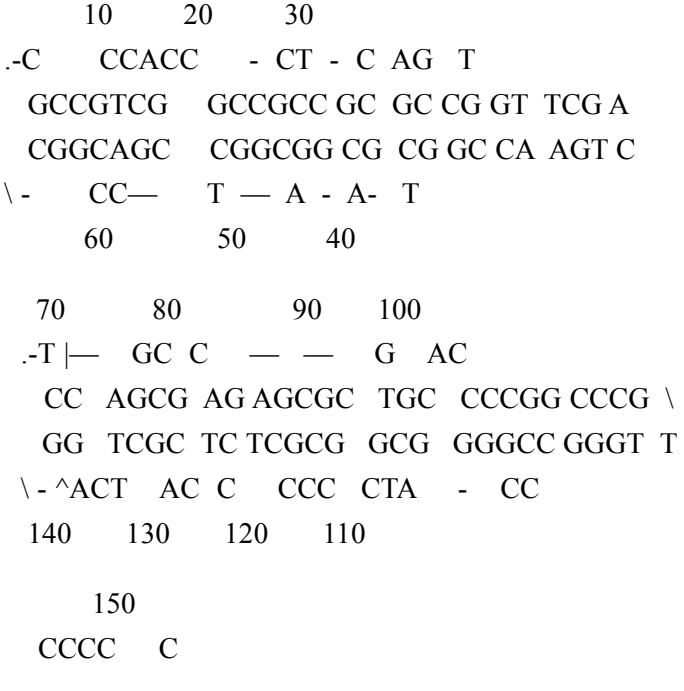




Structure 6

Folding bases 1 to 161 of Chr5:26236024-26236184

Initial dG = -71.0

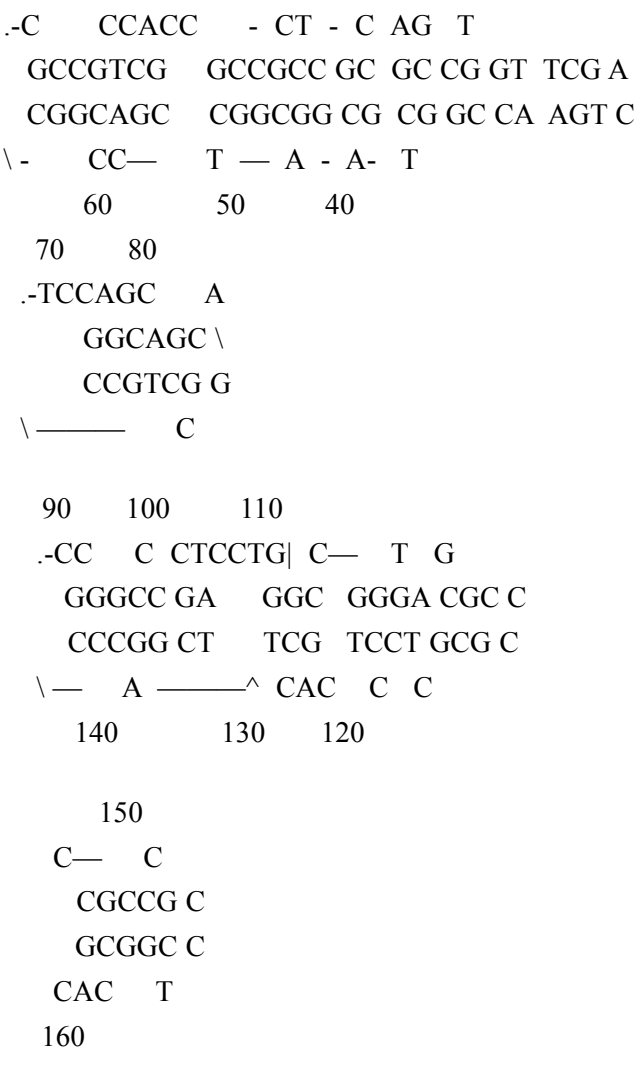


Structure 3……..Structure 6

Structure 7

Folding bases 1 to 161 of Chr5:26236024-26236184

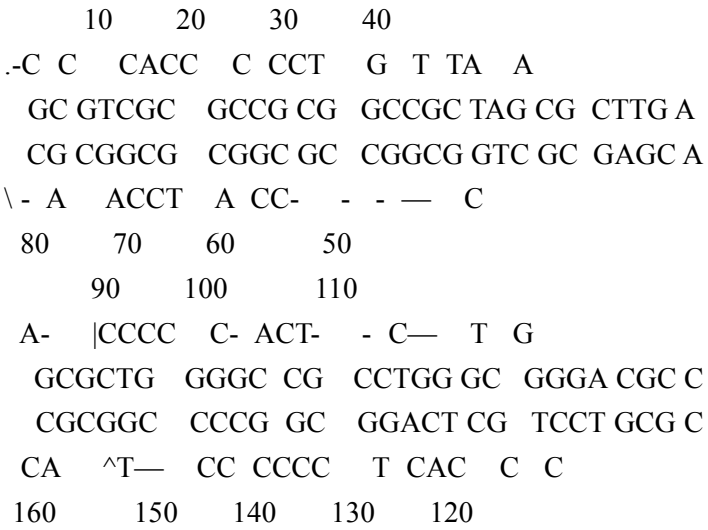
Initial dG = -70.6



Structure 8

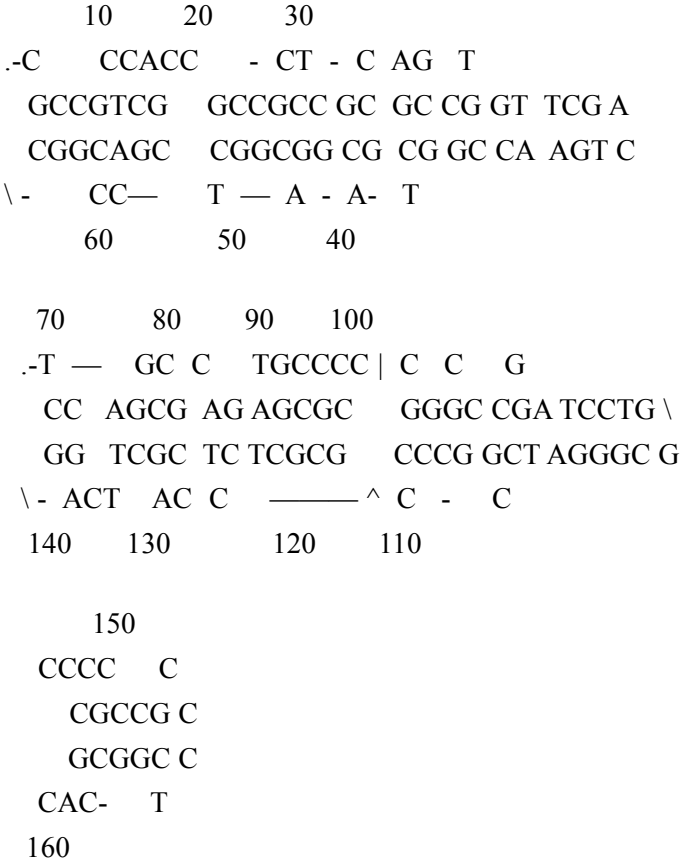
Folding bases 1 to 161 of Chr5:26236024-26236184

Initial dG = -69.6



Structure 10

Folding bases 1 to 161 of Chr5:26236024-26236184
Initial dG = -69.0



4.4 Step4

This programme discovers the good miRNA structures from the above predicted structures where the miRNA was found.

Input:MIRNA.out (i.e. structures found in step 3)

Perl CheckMirs.pl()

Output file: Good_Structure.txt

Chr5:26236024-26236184

5p Structure 8

Folding bases 1 to 161 of Chr5:26236024-26236184

Initial dG = -69.6

```
         10        20        30        40
.-C  C     CACC    C  CCT     G   T  TA    A
  GC GTCGC    GCCG CG   GCCGC TAG CG  CTTG A
  CG CGGCG    CGGC GC   CGGCG GTC GC  GAGC A
\ -  A     ACCT    A  CC-     -   -  —    C
 80        70        60          50
        90        100          110
 A-     |CCCC    C-  ACT-     -  C—    T   G
   GCGCTG    GGGC  CG    CCTGG GC   GGGA CGC C
   CGCGGC    CCCG  GC    GGACT CG   TCCT GCG C
 CA     ^T—    CC  CCCC     T  CAC    C   C
160         150       140       130       120
End Record 0
```

Good Structure Count: 1

5. CONCLUSION

In the present study, we developed an algorithm for prediction of miRNA in plants using shell scripting under Linux environment. Based on this, an automation software was developed which generates RNA secondary structure, RNA structure in XML format and RNA structure in pictorial view and finally predicts the good miRNA structure. This algorithm takes less execution time and memory when compared with earlier algorithms.

REFERENCES

[1] Lee RC, Feinbaum RL, Ambros V. The *C. elegans* heterochronic gene lin-4 encodes small RNAs will antisense complementary to lin-14. Cell 1993;75:843–54.

[2] Reinhart BJ, Slack FJ, Basson M, Beltinger JC, Pasqvinetti AE, Rougvie AE, et al. The 21 nucleotide let-7 RNA regulates developmental timing in *Caenorhabditiselegans*. Nature 2000;403:901–6.

[3] Bushati N, Cohen SM. MicroRNA functions. Annu Rev Cell Dev Biol2007;23:175–205.

[4] Lagos-Quintana M, Rauhut R, Lendeckel W, Tuschl T. Identification of novel genes coding for small expressed RNAs. Science 2001;294:853–8.

[5] Lim LP, Glasner ME, Yekta S, Burge CB, Bartel DP. Vertebrate microRNAgenes. Science 2003;299:15–40.

[6] Lagos-Quintana M, Rauhut R, Meyer J, Borkhardt A, Tuschl T. New microRNAs form mouse and human. RNA 2003;9:175–9.

[7] Dostie J, Mourelatos Z, Yang M, Sharma A, Dreyfuss G. Numerous microRNPsinneuronal cell containing novel microRNA. RNA 2003;9:180–6.

[8] Jones-Rhoades M, Bartel DP, Bartel B. MicroRNAs and their regulatory roles in plants. Annu Rev Plant Biol 2006;57:19–53.

[9] Bartel D. MicroRNAs: genomics, biogenesis, mechanism, and function. Cell 2004;16:281–97.

[10] Lai EC, Tomancak P, Williams RW, Rubin GM. Computational identification of *Drosophila* microRNA genes. Genome Biol 2003;4:R42.

[11] Ritchie W, Legendre M, Gautheret D.RNA stem–loops: to be or not to be cleaved by RNAs III. RNA 2007;13:457–62.

[12] Xue C, Li F, He T, Liu G, Li Y, Zhang X. Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine. BMC Bioinformatics 2005;6:310.

[13] Jiang P, Wu H, Wang W, Ma W, Sun X, Lu Z. MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features. Nucleic Acids Res 2007;35:W339–44.

[14] Xuan P, Guo MZ, Wang J, Wang CY, Liu XY, Liu Y. Genetic algorithm-based efficient feature selection for classification of pre-miRNAs. Genet Mol Res 2011;10(2):588–603.

[15] Molnar A, Schwach F, Studholme DJ, Thuenemann EC, Baulcombe DC. MiRNAs control gene expression in the single-cell alga *Chlamydomonasreinhardtii*. Nature 2007;447:1126–9.

[16] Brennecke J, Hipfner DR, Stark A, Russell RB, Cohen SM. Bantamencodes a developmentally regulated microRNA that controls cell proliferation and regulates the pro-apoptotic gene *hid* in *Drosophila*. Cell 2003;113:25–36.

[17] Xu P, Vernooy SY, Guo M, Hay BA. The*Drosophila*microRNA*miR-14*suppresses cell death and is required for normal fat metabolism. Curr Biol 2003;13:790–5.

[18] Chen CZ, Li L, Lodish HF, Bartel DP. MicroRNAs modulate hematopoietic lineage differentiation. Science 2004;303:83–6.

[19] Krichevsky AM, King KS, Donahue CP, Khrapko K, Kosik KS.A microRNA array reveals extensive regulation of microRNAs during brain development. RNA2003;9(10):1274–81.

[20] Calin GA, Sevignani C, Dumitru CD, Hyslop T, Noch E, Yendamuri S, Shimizu M, Rattan S, Bullrich F, Negrini M, Croce CM. HumanmicroRNA genes are frequently located at fragile sites and genomic regions involved in cancers. Proc Natl Acad Sci U S A 2004;101(9):2999–3004.

[21] Pfeffer S, Zavolan M, Grässer FA, Chien M, Russo JJ, Ju J, John B, Enright AJ, Marks D, Sander C, Tuschl T. Identification of virus-encoded microRNAs. Science2004;304(5671):734–6.