Journal of Root Crops, 2013, Vol. 39 No. 1, pp. 22-27
Indian Society for Root Crops
ISSN 0378-2409

An Algorithm for Plant MicroRNA Target Prediction

C. S. Abhinand, V. S. Santhosh Mithra and J. Sreekumar

Central Tuber Crops Research Institute, Sreekariyam, Thiruvananthapuram 695 017, Kerala, India
Corresponding author: J. Sreekumar, e-mail: sreejyothi_in@yahoo.com
Received: 23 April 2013; Accepted: 30 June 2013

Abstract

MicroRNAs (miRNAs) are RNAs of ~ 24-nucleotide in length which by binding to the 3' untranslated region (3'-UTR) of the target mRNA, bearing complementary target sequences or degrading mRNA by cleaving at single site, cause translational suppression and thereby down–regulate gene expression. The study of the relationship between miRNAs and their target mRNAs is now an attractive area in bioinformatics. Predicting miRNAs, which target mRNAs using experimental methods is a challenging task as it is time consuming and costly. In the present study, for predicting the target sequences in mRNAs, a computational method namely "miRNA target plot" was developed using an efficient R program involving data input, target prediction and plotting. The mature miRNA sequence and specific mRNA sequence information were entered and the sequence information was read using seqinR package. Input data sequences were further processed in two steps. In the first step, the user input miRNA sequence as fasta format in 5'-3' direction was reversed in 3'- 5'direction using function rev.comp ( ) in R package sequinR and the complement of this sequence was used for finding the optimal match based on sequence similarity. In the second step, the program cuts the mRNA sequence from the first position onwards till the end of mRNA sequence and equals the length akin to the length of miRNA. All the possible alignments between each miRNA-mRNA pair was determined using dynamic programming and the scores were calculated. The positive scores were filtered out and the optimal target sequence was found. The miRNA and mRNA segment alignment scores and mRNA segment positions were also plotted in a scatter plot. The computational method is equally effective in predicting target mRNAs compared with other existing tools like TAPIR, psRNAT which has been verified with sequences from StarBase database. The miRNA target plot can be used for precisely predicting target mRNAs for miRNAs.

Key words: miRNA, miRNA target, dynamic programming

Introduction

MicroRNAs (miRNAs) are small RNA molecules which have recently gained widespread attention as critical regulators in complex gene regulatory networks in eukaryotes. The microRNAs, by binding to the 3' untranslated region (3'-UTR) of the target mRNA bearing complementary target sequences or degrading mRNA by cleaving at single site cause translational suppression and thereby down–regulate gene expression (Ambros, 2001; 2004). MicroRNAs have been reported to control a wide range of biological processes such as hematopoiesis, neurogenesis, cell cycle control, and oncogenesis, indicating that they are core elements of the complete gene regulatory network, together with transcription factors (Bushati and Cohen, 2007; Ying et al., 2008). These small RNAs, processed from non-coding regions of the genome into ~ 24 nucleotide long single stranded RNAs, have been shown to regulate translation of messenger RNA (mRNA) by binding to it and effecting target cleavage or destabilize causing translational block depending on the extent of sequence complementarity with the target mRNA. MicroRNAs were discovered in 1993, in a genetic screening for mutants that disrupted the timing of post-embryonic development in the nematode *Caenorhabditis elegans*. When the first miRNA, let7, was discovered and found

to be highly conserved in eukaryotes, it led to a surge in discovery of new miRNAs in a number of organisms including humans. Most known miRNAs are very well conserved in closely related species, while some can be found across very large taxonomic groups, notably let7 of *C. elegans* (Lee et al., 1993; Reinhart et al., 2000).

miRNA biogenesis and action

The miRNA genes are frequently expressed individually, but may exist in clusters of 2–7 genes as coding regions with small intervening sequences (Lau et al., 2001; Lee et al., 2002; Baskerville and Bartel, 2005). The miRNA biogenesis in animals is a two-step process. In the first step, miRNA is transcribed as longer RNA molecule called pri-miRNA (Lai, 2003; Cai et al., 2004; Borchert et al., 2006) and the pri-miRNA is processed in the nucleus itself into hairpin RNA of 60 to 120 nucleotides by a protein complex consisting of the ribonuclease Drosha and an RNA binding protein Pasha. This hairpin RNA, known as pre-miRNA, is transported to the cytoplasm via exportin-5 dependent mechanism as the second step. It is digested there by a dsRNA specific ribonuclease called Dicer to form the mature miRNA (Hutvagner and Zamore, 2002; Zhang et al., 2004). Mature miRNA is bound by a complex, similar to the RNA induced silencing complex (RISC) that participates in RNA interference (RNAi). The mature miRNA makes base pairing with mRNA wherever complementarities exist between them.

The way miRNA and their target mRNAs interact are different in animals and plants in certain aspects. In plants, miRNA exhibits perfect or nearly perfect base pairing with the target mRNA but in the case of animals, the pairing is rather imperfect (Bartel and Barte, 2003). This makes the miRNA target identification in animals more complex compared to that in plants. Also miRNAs in plants bind to their target mRNAs within coding regions cleaving at single sites whereas most of the miRNA binding sites in animals are in the 32 un-translated regions (UTR) (Bartel and Barte, 2003). This results in mRNA target degradation in plants and destabilization in animals (Kloosterman et al., 2004; Lytle et al., 2007). Single mRNA can contain multiple miRNA target sites for different miRNAs or for the same miRNA. It is also known that miRNAs are highly conserved among different species. In addition to the conserved miRNAs, there are lots of non conserved species specific miRNAs that may control the specific characteristics that are unique to those species (Lai, 2004).

miRNA target prediction

The prediction methods are mainly classified into experimental method and sequence based method (Ander Muniategui et al., 2012). The most extensively used experimental technique for determining miRNA targets is the transfection of mimic miRNAs or miRNA inhibitors. The effects on the expression levels of the mRNAs and proteins are measured using transcriptomic and proteomic tools (qRT-PCR, microarrays, RNA-seq, western blot, SILAC, 2D-DIGE). However, this technique does not distinguish indirect and direct interactions between miRNA and mRNA. Other direct methods for miRNA target prediction are based on the immunoprecipitation of RISC complexes such us Argonaute bound miRNA–mRNA molecules. Each experimental technique has its own reliability. Due to this, combining different experimental tools is a good method to ensure the authenticity of a miRNA target (Ander Muniategui et al., 2012).

Despite the wide range of available experimental tools for miRNA target prediction and validation, the lack of high-throughput and low-cost methods have enforced the development of computational method. These methods are based on experimentally verified thumb rules for miRNA targeting:

(i) Sequence complementarity between the 3'-UTR of the mRNAs and the 'seed region' (region of about 6-8 nucleotides in length at the 5' end of an animal miRNA) of the miRNA.

(ii) Possible functional target sites along the coding sequence and 5'-UTR of the mRNA.

(iii) Conservation of some of the miRNA target sites between related species.

(iv) The target site accessibility due to the RNA secondary structure *(*i.e., free energy costs to unfold the mRNA secondary structure surrounding the target site and free energy of the miRNA-target pairing).

Although the methods that use these rules are far from perfect, the putative lists of target mRNAs generated by

computational methods entangle a considerable reduction of experimental work as they significantly reduce the number of interactions that must undergo validation (Ander Muniategui et al., 2012; Sungroh Yoon and Giovanni De Micheli, 2006).

The computational methods for miRNA target prediction can be broadly classified as

1. Complementarity searching based method

2. Thermodynamic based method

3. Machine learning method

Complementarity searching based method identifies initial potential targets using complementarity searching algorithms and then improves them by using other features like thermodynamics, binding site structure and conservation. Stark and coworkers initially implemented this strategy for predicting miRNA targets in *Drosophila melanogaster*. The miRanda, TargetScan and PicTar were developed based on this strategy (Stark et al., 2003). Thermodynamic based method uses the favorable thermodynamic structure as an initial indicator and then uses other properties of miRNA-mRNA interaction to filter miRNA targets. The DIANA-microT and RNAHybrid falls under this category (Mendes et al., 2009). The 3'UTRs have some common motifs of short length, some of these are complementary to the seed region of known miRNA, and others might be the target of unknown miRNAs. This method of analyzing whole genome to predict miRNA targets can only be applied to predict conserved targets. Machine learning methods like SVM, HMM and ANN are also used to predict miRNA targets but the performance of machine learning method is affected by shortage of training data. The miTarget classifier is an example for SVM based predictor (Kim et al., 2006).

The objective of the present study was to develop a computational method for predicting miRNA target mRNA provided the miRNA and mRNA sequences of the plant are known. It has been established from the experimentally validated plant miRNAs targets that miRNA and their target mRNA have high complementarity between them. Hence this approach mostly concentrated on the sequence similarity.

Materials and Methods

R, an open source programming environment for statistical computing was used to implement the prediction algorithm as shown in the Fig.1.

For the implementation of the above work plan in R suit, the algorithm was divided into six different parts and each part was listed individually.

Input data

The mature miRNA sequence and specific mRNA sequence information of the intended organism were entered. The miRNA

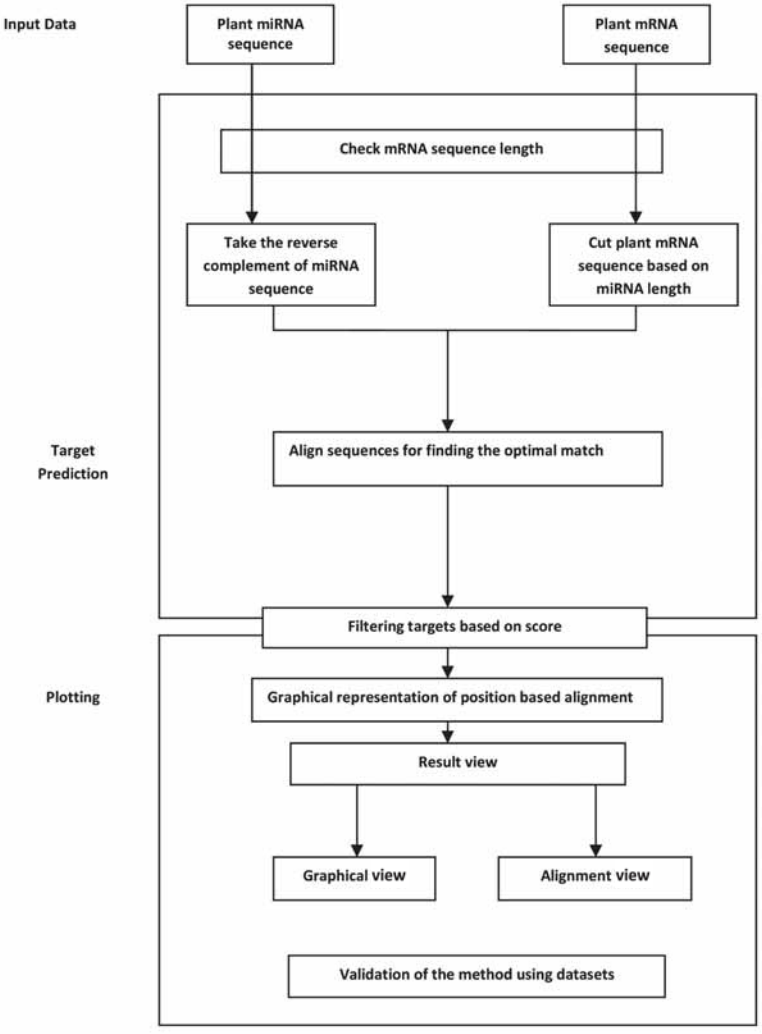


Fig. 1. The schematic representation of the work

sequence was in 5'-3' direction and length in the range of 19 - 28 nucleotides. The mRNA sequence length had no limitations and it was also in 5'-3' direction. Both sequence information were stored as fasta sequence format and saved as fasta files.

Sequence information was read using seqinR package, which is a library of utilities to retrieve and analyze biological sequences. Input data sequences were further processed in two steps. In the first step, the user input miRNA sequence as fasta format in 5'-3' direction was reversed in 3'- 5'direction using function rev.comp ( ) in R package sequinR and the complement of this sequence was used for finding the optimal match based on sequence similarity. In the second step, the program cut the mRNA sequence from the first position onwards till the end of mRNA sequence and equaled the length with the length of miRNA. This helped to find out all the possible alignments between each miRNA-mRNA pair. The process started from the first nucleotide position and continued till the end of mRNA sequence.

Align miRNA and mRNA

The miRNA sequence was first aligned with each mRNA sequence segments using dynamic programming. Needleman Wunch algorithm for global alignment was used to find out the optimal match target sequence. The nucleotide SubstitutionMatrix ( ) function in the Biostrings package in R was used to obtain alignment. Here gaps were allowed between the miRNA-mRNA sequences, but mismatches were preferred to gaps by giving a higher penalty for gaps. To reduce the number of predicted false positive targets, the program limited the number of mismatches, and instead were assigned suitable scores; if the alignment score was higher, the target sequence had more complementarity with the input miRNA sequence. To attain this, a scoring scheme with sequence match for a score of (1), a score of (-1) for each mismatch, for gap opening penalty a score of (-8) and a gap extension penalty (-2) were followed. To align the sequences, a scoring matrix sigma was created based on these scores. In this sequence alignment process, the program represented each miRNA segments as "pat" and the mRNA sequence as "sub" means pattern and subject of the alignment. Each mRNA position held the variable "i" and the scores for all alignment were stored at the variable "scores".

Filtering with positive scores

The program stored all alignment scores into a variable "scores". The positive scores were filtered out from the scores which helped to find out the optimal target sequence. After the first filtering the program retrieved the sequence segment that had maximum score of the alignment. The sequences that had alignment score near to the maximum score were also considered for giving a probabilistic output.

Constructing miRNA target plot

For each position over the length of mRNA, the alignment scores were taken from the global alignment and were plotted in a scatter plot. The X axis of the graph represented the position and the Y axis showed the alignment scores. The plot showed the whole alignment between miRNA and mRNA in plants. It also helped to find out the maximum score, i.e. the putative target region of mRNA. The graphical output showed the interaction between miRNA and mRNA sequence in plants and the highly probable alternate targets.

Predicting miRNA targets

In this step, the targets were selected based on the similarity scores and graphical and text outputs were displayed. The text output showed the alignment between miRNAs and mRNA and its corresponding alignment scores while the graphical output showed the whole alignment between miRNA and mRNA as a scattered one. The target segment with maximum alignment score gave the higher priority.

Validation of the program

Datasets from the StarBase database was used to validate the program. This is a public platform for decoding microRNA-target and protein-RNA interaction maps from CLIP-Seq and degradome sequencing data.

Results and Discussion

The output report consisted of two parts. The first part consisted of sequence alignment result, including the pattern and subject. The next part was a graphical output that helped to determine the whole alignment between miRNA and mRNA in plants. Information shown for each predicted target included target site position and corresponding alignment score (Fig. 2b). The target was indicated if it had high complementarity to miRNA. It

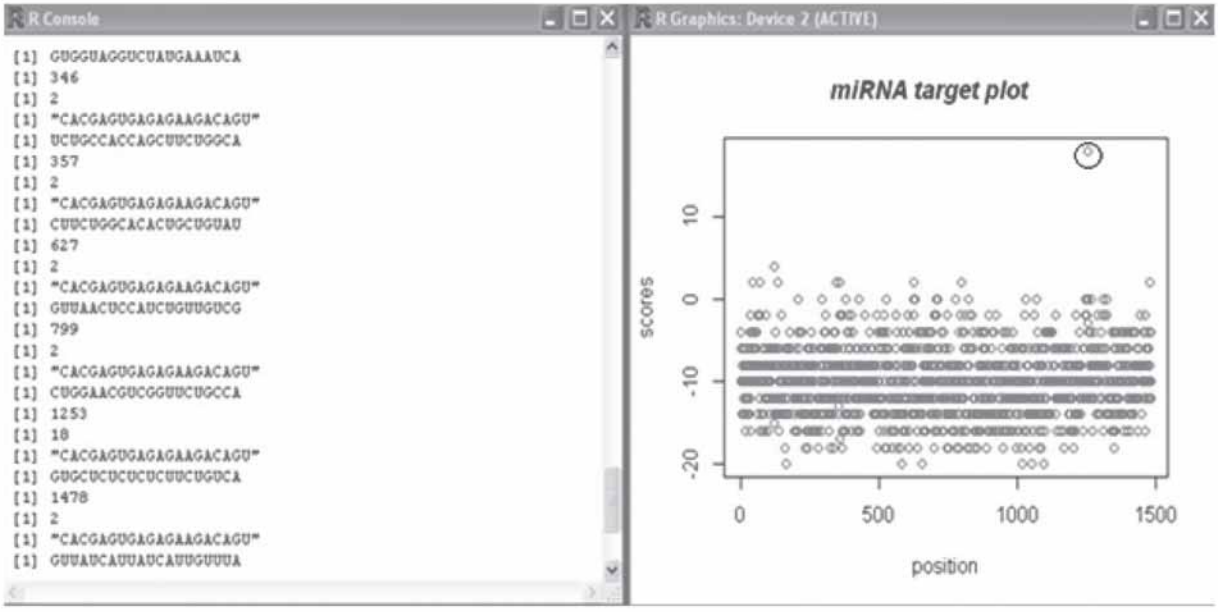
also included highly probable alternate targets which had high alignment score. For example in the output text (Fig. 2a) results show a score of 18 in the location 1253 while other locations have 2 as score. So we can expect a target location at 1253 with higher probability as the target.

For this computational method, *Arabidopsis thaliana* miRNAs downloaded from miRBase (www.mirbase.org) were used as one of the query for the program to predict target mRNAs and the mRNA data were retrieved from TAIR (www.arabidopsis.org) database. All the reported potential target sequences (locations) of miRNAs were successfully detected by this method. Most of the interactions between plant miRNA and mRNA documented in StarBase were identified by this method. Compared to StarBase tool the miRNA target plot method could identify target sites of mRNA and a plot of miRNA-mRNA interaction was drawn. The scatter plot gave a clear picture of the locations of the miRNA targets. Also the method did not miss any short target of a plant mRNA having high probability of similarity. After testing this method with a set of *Arabidopsis* genome data, the test was repeated for existing tools like TAPIR and psRNAT, with the same data set. The developed method gave its best performance when compared with the result of other tools using the same data set.

The method developed in the present study could predict 92% of the targets predicted and recorded by StarBase while the remaining showed some positional changes near to the original. The area under Receiver Operating Characteristic (ROC) curve was computed using the true targets (considered the targets in StarBase) and the targets predicted by this method was 0.96, which showed the high target prediction capability of the method. The reason for identifying targets by the miRNA target plot method was that most plant miRNAs show near-perfect or perfect complementarity to their targets.

Conclusion

Even though many computational tools have been developed for miRNA target prediction, few have been designed exclusively to find plant miRNA targets. In the present study a miRNA target prediction method has been developed based on dynamic programming in R environment for statistical computing. The method was found equally effective in predicting plant miRNA targets in mRNA sequences compared with other existing tools like TAPIR, psRNAT which have been verified with sequences data available from StarBase database. The scores obtained for each targeted location plotted in the scatter diagram helped in selecting location with a high probability of an interacting target.



a b

Fig. 2a. Text display of mRNA target identified by filtering

Fig. 2b. The miRNA target plot

References

Ambros, V. 2001. MicroRNAs: tiny regulators with great potential. *Cell*, **107**: 823–826.

Ambros, V. 2004. The functions of animal microRNAs. *Nature*, **431**:350–355.

Ander Muniategui., Jon Pey., Francisco Planes and Angel Rubio. 2012. Joint analysis of miRNA and mRNA expression data. *Briefings in Bioinformatics*.

Bartel, B. and Barte, D. P. 2003. MicroRNAs: at the root of plant development? *Pl. Physiol*, **132**: 709–717.

Baskerville, S. and Bartel, D. P. 2005. Microarray proling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes. *RNA*, **11**: 241–247.

Borchert, G.M., Lanier, W. and Davidson, B. L. 2006. RNA polymerase III transcribes human microRNAs. *Nat. Struct. Mol. Biol*., **13**: 1097–1101.

Bushati, N. and Cohen, S.M. 2007. MicroRNA functions. *Ann. Rev. Cell Dev. Biol*., **23**: 175- 205.

Cai, X., Hagedorn, C.H. and Cullen, B. R. 2004. Human microRNAs are processed from capped, polyadenylated transcripts that can also function as mRNAs. *RNA*, **10**: 1957–1966.

Hutvagner, G. and Zamore, P. D. 2002. A microRNA in a multiple turnover RNAi enzyme Complex. *Sci*., **297**: 2056–2060.

Kim, S. K., Nam, J. W., Rhee, J. K., Lee, W. J. and Zhang, B. T. 2006. miTarget: microRNA target gene prediction using a support vector machine. *BMC Bioinformatics*, **7:** 411.

Kloosterman,W. P., Wienholds, E., Ketting, R. F. and Plasterk, R. H. A. 2004. Substrate requirements for let-7 function in the developing zebra fish embryo. *Nucleic Acids Res*., **32**: 6284–6291.

Lai, E. C. 2003. microRNAs: runts of the genome assert themselves. *Curr. Biol*., **13**: R 925– R 936.

Lai, E.C. 2004. Predicting and validating microRNA targets. *Genome Biol*., **5**:115.

Lau, N. C., Lim, L.P., Weinstein, E. G. and Bartel, D. P. 2001. An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans*. *Sci*., **294**: 858–862.

Lee, R. C., Feinbaum, R. L. and Ambros, V. 1993. The *C. elegans* heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*., **75**: 843 – 854.

Lee, Y., Jeon, K., Lee, J.T., Kim, S. and Kim, V.N. 2002 . MicroRNA maturation: stepwise processing and subcellular localization. *EMBO J*., **21**: 4663–4670.

Lytle, J. R., Yario, T. A. and Steitz, J. A . 2007. Target mRNAs are repressed as efficiently by microRNA-binding sites in the 5' UTR as in the 3' UTR. *Proc. Natl Acad. Sci*. USA, **104**: 9667–9672.

Mendes, N. D., Freitas, A. T. and Sagot, M. F. 2009.Current tools for the identification of miRNA genes and their targets. *Nucleic Acids Res*., **37**(8): 2419– 2433.

Reinhart, B. J., Slack, F. J. and Basson, M. 2000. The 21-nucleotide let-7 RNA regulates developmental timing in *Caenorhabditis elegans*. *Nature*, **403**: 901– 906.

Stark, A., Brennecke, J., Russell, R.B. and Cohen, S.M. 2003. Identification of *Drosophila* MicroRNA targets. *PLoS Biol*., **1** (3): 397-409.

Sungroh Yoon and Giovanni De Micheli. 2006. Computational identification of microRNAs and their targets. *Birth Defects Research* (Part C), **78**: 118-128.

Ying, S.Y., Chang, D. C. and Lin, S. L. 2008. The MicroRNA (miRNA): overview of the RNA genes that modulate gene function. *Mol. Biotechnol*., **38**(3): 257-268.

Zhang, H., Kolb, F. A., Jaskiewicz, L., Westhof, E. and Filipowicz, W. 2004. Single processing center models for human Dicer and bacterial RNase III. *Cell*, **118**: 57–68.