Computational analysis of miRNA targets in plants: current status and challenges

Xinbin Dai, Zhaohong Zhuang and Patrick Xuechun Zhao

Submitted: 28th June 2010; Received (in revised form): 21st August 2010

Abstract

Plant microRNAs (miRNA) target recognition mechanism was once thought to be simple and straightforward, i.e. through perfect reverse complementary matching; therefore, very few target prediction tools and algorithms were developed for plants as compared to those for animals. However, the discovery of transcription suppression and the more recent observation of widespread translational regulation by miRNAs highlight the enormous diversity and complexity of gene regulation in plant systems. This, in turn, necessitates the need for advanced computational tools/algorithms for comprehensive miRNA target analysis to help understand miRNA regulatory mechanisms. Yet, advanced/comprehensive plant miRNA target analysis tools are still lacking despite the desirability and importance of such tools, especially the ability of predicting translational inhibition and integrating transcriptome data. This review focuses on recent progress in plant miRNA target recognition mechanism, principles of target prediction based on these understandings, comparison of current prediction tools and algorithms for plant miRNA target analysis and the outlook for future directions in the development of plant miRNA target tools and algorithms.

Keywords: plant miRNA; miRNA target prediction; translational inhibition; post-transcriptional regulation

MECHANISM OF miRNA TARGET RECOGNITION IN PLANTS

MicroRNAs (miRNAs) are 20- to 24-nt endogenous non-protein-coding RNAs that regulate gene expression at post-transcriptional or translational level. In eukaryotes, miRNAs are able to direct the RNA-induced silencing complexes (RISCs) to their complementary target sites in mRNAs after binding to ARGONAUTE (AGO) proteins. The RISCs regulate target gene expression via two different mechanisms: (i) translational repression, sometimes coupled with accelerated mRNA decay [1, 2] and (ii) RISC-catalyzed endonucleolytic mRNA cleavage [3]. The former directly affects protein translation efficiency and the latter reduces mRNA expression. The mRNA degradation, also called post-transcriptional regulation, was considered as a major regulatory mechanism in plant systems while translational inhibition by miRNAs has been thought to be a major mechanism for animal systems.

Recent evidences suggest that plant miRNA-guided silencing has a widespread translational inhibitory component, which is genetically separable from endonucleolytic cleavage [4]. For example,
Dugas and Bartel [5] reported that the altered miR398 complementarity sites do not affect mRNA accumulation of its target genes, CSD1 and CSD2, whereas the encoded protein accumulation of both the genes remain sensitive to miR398 levels in Arabidopsis thaliana, suggesting that miR398 acts as a translational repressor when target site complementarity is reduced. In the same study, Dugas and Bartel reported that miR834, which is one of the non-conserved ('young') miRNAs, primarily decreases its target gene CIP4's protein level rather than mRNA level [5]. The miRNA translational inhibition mechanism also requires the activity of the microtubule-severing enzyme katanin, implicating cytoskeleton dynamics in miRNA action, as recently suggested from animal studies.

On the other hand, recent studies also suggest that a number of miRNAs in animal systems negatively regulate the expression of protein-coding genes by the cleavage of the mRNAs [6]. These discoveries together indicate that both plants and animals probably share similar mechanisms in the regulation of target genes’ expression.

The crystal structure of the Argonaute protein from Pyrococcus furiosus has been resolved at 2.25 Å resolution. The 3D structure reveals a Piwi Argonaute Zwille (PAZ) domain that forms a groove along with N-terminal, PIWI domain and middle domain. Song et al. [7] proposed a miRNA–mRNA interaction model, in which the 3’-end of small RNA binds to one end of PAZ cleft and the 5’-end binds near the other end of the PAZ cleft. The mRNA enters in the groove between the N-terminal and PAZ domains, and comes out between the PAZ and middle domain. Since the groove is narrower on middle domain side, the 5’-end small RNA and 3’-end mRNA on this side are closer and paired by nucleotides, which will bind mRNA to the whole RISC complex (Figure 1A). This model emphasizes the importance of the first 8 bp in small RNA/mRNA recognition. Systematic miRNA target site mutagenesis in human and Arabidopsis also showed that the base pairing at the 5’-end of miRNA is important for target recognition [8, 9]. Phylogenetic analysis of miRNAs across species also validates that the first 8 nt are conserved and often function as determinant of target recognition [10]. Therefore, the seed region (the first 6–8 bp of miRNA) has been emphasized in miRNA target analysis in both plants and animals.

**PRINCIPLES OF PLANT miRNA TARGET PREDICTION**

**Complementarity between miRNA and target site**

In plants, the complementarity between miRNA and target site determines the stability of miRNA–mRNA duplex and therefore has been utilized as a key feature for target gene analysis.

Systematic mutagenesis studies highlighted the seed regions for miRNA targeting in Arabidopsis.
In general, a scoring schema that expects a perfect match or only allows one GU wobble between seed region of miRNA and target site of mRNA has been widely employed in published plant miRNA prediction tools. Such criterion was further validated by experiments. However, some exceptions to this criterion have been reported. For example, miR-398a regulates its target gene CSD2 at post-transcriptional level in Arabidopsis, though the seed region contains a bulge and GU wobble [11]. A similar phenomenon was noticed between miR396a and its target gene GRL7, GRL8 and GRL9. Thus, a stringent seed region matching rule may miss potential target sequences for some miRNAs.

Besides seed region, miRNA nucleotides 16 and 19 are thought to be important for target recognition in Arabidopsis. Tcp4-soj18 mutation at position 19 of miR319 or position 16 of mRNA target site significantly affects their pairing, although these changes cause only a small difference in the calculated interaction free energy [12]. These exceptions imply the complexity of recognition mechanism between miRNA and target mRNA, thereof demanding improved methods over traditional prediction tools that only applies simple matching rule between miRNA and target mRNA.

### Binding site evolutionary conservation

The conservation of target sequence is another criterion commonly applied to filter miRNA target candidates. In some target prediction algorithms, the existence of conserved potential miRNA target(s) in orthologous sequences in closely related species have been used to re-enforce the prediction [13]. One of the challenges with such an analysis is that the expression pattern of miRNA targets may vary in different tissues and developmental stages, and target mRNA sequences may not be detected due to their low expression level in specific organisms [14], leading to false negative predictions from conservative analysis.

### Multiplicity of target sites

The existence of multiple potential target sites is frequently viewed as a necessity for efficient translational inhibition in animals. In reported plant miRNA target prediction tools, the importance of the target site multiplicity was generally underestimated while highlighting the miRNA’s ability to closely bind to/perfectly match the target site. However, increasing reports on translational inhibition of the miRNA targets in Arabidopsis have emphasized the importance of the effect of multiplicity of target sites. For example, Axtell et al. [15] reported dual target sites on AtTAS3 which is the target gene of miR390. AtTAS3 is the phasing precursor of ta-siRNA TAS3 and the biogenesis of the ta-siRNA shares similar target recognition mechanism with miRNA [15]. In this case, two target sites (two hits) represent two cleavage events, which constitute a stronger signal for degradation of target mRNA.

### mRNA degradation or translational inhibition

A major update on plant miRNA-guided silencing is the discovery of widespread translational inhibitory component that is genetically separable from endonucleolytic cleavage. Broderson et al. concluded that central mismatches in miRNA-target mRNA duplex lead to translational repression because it prevents slicing; while central matches in miRNA-target mRNA duplex tend to cleave target mRNA and exclude translational repression regardless of a few mismatches in other regions [4] (Figure 1B). In addition, the translational target sites seem to be located in coding region rather than in 3′-UTR region as in animals. These differences make target prediction tools designed for animal systems unsuitable for plant miRNA target prediction.

This suggests that many ‘non-functional’ miRNAs, especially non-conserved ‘orphan’ miRNAs, need to be re-identified since the experimental methods, such as 5′-RACE PCR method, would fail to detect the change of target genes at protein level [4]. To address this issue, novel target prediction algorithm enabling translational analysis is desired by the plant miRNA research community.

### Target site accessibility

Physicochemistry of miRNA–target interactions also needs to be taken into account for miRNA target recognition. The free energy (or thermo stability) of interaction between a miRNA and its target is generally not a very good predictor, whereas the miRNA target site accessibility has been proved to be one of the effective structural features for the prediction of target site functionality [16–18], i.e. an effective target site on mRNA is expected to have less mRNA secondary structural blockage,
which affects the miRNA’s access to the target site. The RNAup program in Vienna package is often used to calculate secondary structure in mRNA target site region [19].

COMMON TARGET PREDICTION TOOLS OR ALGORITHMS FOR PLANT miRNAs AND TA-siRNAs

A number of algorithms and programs have been developed to analyze complementarity between miRNAs and their target sites. One of the earliest programs used for searching complementary target sites is PatScan [20], which has been successfully applied in studying miRNAs in rice and Arabidopsis [11, 21, 22]. By inputting user-defined pattern sets, PatScan is able to search the complementary matching sequences that allow mismatch, GU wobble and bulged nucleotide (insertion/deletion) between miRNA and target site. However, it is a challenge to generate comprehensive pattern sets for large-scale systematical miRNA target analysis. And, extra skill or programming work is required in order to effectively use PatScan, for example to consider individual weighting on each match/mismatch between seed and central regions. Fahlgren and Carrington [23] described a position-dependent scoring system for predicting plant miRNA targets. The scoring system is based on FASTA program and wrapped by a Perl script, which is able to consider mismatch, single-nucleotide bulges or gaps. Seed region and central region are individually scored, with heavily penalized mismatches on position 2 through 13 [23].

Zhang [25] introduced a web tool, miRU, for plant miRNA target prediction. The miRU was designed on the basis of Smith–Waterman algorithm [24]. Smith–Waterman algorithm is a well-known algorithm for performing local sequence alignment, which returns an optimal alignment between miRNA and target sequence, regardless of the length of miRNA. Single miRNA sequence can be submitted to search potential target sequences in cDNA libraries hosted on the miRU server; however, it lacks the ability to accept sequences in batch and custom target candidates. The miRU also outputs target sequence conservation information between species [25]. Secondly, the Smith–Waterman algorithm is very computationally demanding in terms of both the computation time and memory, which makes it suitable for only small datasets. In addition, this dynamic programming algorithm only returns the optimal alignment and therefore, is unable to calculate the multiplicity of target sites for each target mRNA.

To overcome these limitations of native Smith–Waterman algorithm, Dai et al. implemented a web-based prediction tool for high-throughput miRNA target prediction, namely psRNATarget (X. Dai and P. Zhao, unpublished data, web server available at http://bioinfo3.noble.org/psRNATarget/). In psRNATarget server, the authors developed a parallel iterative Smith–Waterman algorithm on the basis of search program [26], which is enhanced by Gotoh’s improvements [27] and Intel SSE2 instruction set and is able to run on distributed Linux clusters. Therefore, the algorithm is capable of analyzing high-throughput small RNA data on genome scale and output multiplicity of target sites. Furthermore, psRNATarget individually scores each match/mismatch in seed region, central region and other 3’-end region with different weights. Tendency of translational inhibition is also reported when central mismatch is detected. Conserved domain information on target candidates is given as a reference for further analysis, such as phylogenetic analysis of target sequences. The psRNATarget server implements three interfaces for users to upload their customized miRNA and target candidate sequences or select published miRNA/cDNA dataset regularly downloaded from public repositories.

Xie et al. [28] developed a BLAST-based target search program, miRNAassist. This program is easy to use and is freely available upon request. Using miRNAassist, the authors identified 67 potential targets for 21 miRNA in Brassica napus. However, the performance of BLAST for miRNA target search is controversial. NCBI BLASTN does not distinguish GU wobble using default scoring matrix, while attempts to use the customized alternate scoring matrices only returns an error message from the program [29].

Another downside of BLASTN is that it may miss potential hits because the BLAST heuristic search algorithm is not suitable to search short sequences, such as 21–24 nt miRNAs. To compare BLAST and Smith–Waterman algorithm, we used 218 published miRNA mature sequences (http://www.mirbase.org/, as of May 2010) and 39 640 cDNA from TAIR9 gene models (http://www.arabidopsis.org/) as testing datasets. NCBI BLASTN search returned
231 cDNA target candidates for 120 published miRNA sequences, while search, a popular Smith-Waterman program, found 863 cDNA target candidates for 207 miRNA mature sequences, if allowing four mismatch for both methods [30] (see Supplementary data). These results indicate that NCBI BLAST can only find less than half of potential candidates compared to the Smith–Waterman program.

The above approaches are mainly based on sequence alignment between miRNA and target mRNA, which are efficient in searching all possible target candidates. However, high false positives may be reported since other factors will also affect the recognition between miRNA and target site. Of these factors, target site accessibility has been proved to be an important determinant for successful target binding process. Li and Zhang [17] developed a pipeline program for Arabidopsis miRNA target gene prediction. In addition to searching complementary sequences and comparing sequence conservation between rice and Arabidopsis, the prediction pipeline also analyzed mRNA secondary structures to evaluate target site accessibility. Using this pipeline, the authors reported 102 miRNA target genes that were classified into 28 transcription factor gene families [17].

Table 1 summarizes the discussed common plant miRNA target prediction tools.

**Table 1:** Comparison of common plant miRNA target prediction programs

<table>
<thead>
<tr>
<th>Allow bulges</th>
<th>PatScan</th>
<th>FASTA based</th>
<th>miRU</th>
<th>psRNATarget</th>
<th>miRNAassist</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weight more in seed region</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Target site multiplicity</td>
<td>No extra work</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Conservation</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Target site accessibility</td>
<td>No extra work</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Detect translational inhibition</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Batch submission</td>
<td>Yes</td>
<td>Yes, but slow</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Designed for biologist?</td>
<td>No</td>
<td>Yes</td>
<td>No, only algorithm</td>
<td>Yes, web</td>
<td>Yes, available</td>
</tr>
<tr>
<td>Availability</td>
<td>Yes, local installation</td>
<td>No, only algorithm description available</td>
<td>Yes, web based</td>
<td>Yes, web based</td>
<td>Yes, available upon request</td>
</tr>
</tbody>
</table>
integrating the degradome profile data, the target sequences that were previously ignored due to low miRNA abundance have been re-identified in Arabidopsis [14, 34]. For example, At1g72230 and At2g44790 genes have been verified as targets of miR408, and At3g14560 is targeted by miR824. It will be very helpful to incorporate ‘omics’ level data in an advanced miRNA target analysis pipeline, though such ‘omics’ level data, especially associated degradome data, are not commonly available for a miRNA project yet.

The recognition of miRNA target is determined by multiple factors, such as complementarity of different regions on miRNA, binding site conservation and target site accessibility. Since the extent of influence of these factors in recognition mechanisms is still unclear to date, current predictive approaches are often based on only parts of these factors; lacking comprehensive consideration of multi-factors may lead to more false positive or negative predictions. Machine learning-based approaches have great potential to integrate these multiple factors through statistical-based learning process using known miRNA–target interaction data and to efficiently incorporate the aforementioned transcriptome and degradome data. Several machine learning-based approaches have already been successfully employed to predict target genes of animal miRNA, e.g. RNA22 [35], miTarget [36] and GenMiR++ [33]. Therefore, integrating such diverse approaches and datasets in a comprehensive manner would substantially improve plant miRNA targets prediction, provided adequate corresponding training datasets are available for plants.

SUPPLEMENTARY DATA
The datasets used to compare BLAST- and Smith–Waterman-based algorithms are available at http://bioinfo.noble.org/manuscript-support/targetreview/.

Key Points
This article covers:
- Basic principles for plant miRNA target recognition and corresponding algorithms.
- New discoveries in plant miRNA target recognition mechanisms, for example, multiple target site recognition, translational inhibition, etc. which demands novel algorithms for plant small RNA regulatory target analysis.
- Future directions on the development of plant miRNA target prediction tools, for example, high-throughput systems to facilitate next-generation sequencing data and machine-learning approaches to comprehensively integrate multi-determinant factors.

Acknowledgements
The authors are grateful to Drs Rakesh Kaundal and James Susamuthu for critical reading of the manuscript and providing valuable comments. The Samuel Roberts Noble Foundation mission statement: to conduct basic biochemical, genetic and genomic research for crop improvement, enhancement of human and animal health and production of novel products in crops. The Noble Foundation is a non-profit research institute.

FUNDING
Oklahoma Center for the Advancement of Science and Technology (OCAST PSB09-32); National Science Foundation (DBI: 0960897); Samuel Roberts Noble Foundation.

References


