psRNATarget: a plant small RNA target analysis server

Xinbin Dai and Patrick Xuechun Zhao*

Plant Biology Division, The Samuel Roberts Noble Foundation, 2510 Sam Noble Parkway, Ardmore, OK 73401, USA

ABSTRACT

Plant endogenous non-coding short small RNAs (20–24 nt), including microRNAs (miRNAs) and a subset of small interfering RNAs (ta-siRNAs), play important role in gene expression regulatory networks (GRNs). For example, many transcription factors and development-related genes have been reported as targets of these regulatory small RNAs. Although a number of miRNA target prediction algorithms and programs have been developed, most of them were designed for animal miRNAs which are significantly different from plant miRNAs in the target recognition process. These differences demand the development of separate plant miRNA (and ta-siRNA) target analysis tool(s). We present psRNATarget, a plant small RNA target analysis server, which features two important analysis functions: (i) reverse complementary matching between small RNA and target transcript using a proven scoring schema, and (ii) target-site accessibility evaluation by calculating unpaired energy (UPE) required to 'open' secondary structure around small RNA's target site on mRNA. The psRNATarget incorporates recent discoveries in plant miRNA target recognition, e.g. it distinguishes translational and post-transcriptional inhibition, and it reports the number of small RNA/target site pairs that may affect small RNA binding activity to target transcript. The psRNATarget server is designed for high-throughput analysis of next-generation data with an efficient distributed computing back-end pipeline that runs on a Linux cluster. The server front-end integrates three simplified user-friendly interfaces to accept user-submitted or preloaded small RNAs and transcript sequences; and outputs a comprehensive list of small RNA/target pairs along with the online tools for batch downloading, key word searching and results sorting. The psRNATarget server is freely available at http://plantgrn.noble.org/psRNATarget/.

INTRODUCTION

Plant endogenous non-coding short small RNAs (20–24 nt), including microRNAs (miRNAs) and a subset of small interfering RNAs (ta-siRNAs), are derived from the cleavage products of double-strand RNAs (ds-RNAs) by DICER-like enzymes (1–4). These regulatory small RNAs (mainly include miRNAs and ta-siRNAs, *sic passim*) negatively regulate gene expression at post-transcriptional level by directing the cleavage of target transcript (mRNA) (5). Several transcription factors and development-related genes have been reported as targets of these regulatory small RNAs, and together they play key roles in gene expression regulatory network controlling plant growth and development (6). These discoveries have aroused wide interest and urgent demand for genome-wide analysis of small RNAs and dissect of their functions, for example, identifying their regulatory target genes in plants.

A number of algorithms and programs have been introduced to search target genes for miRNAs (7–9). However, most of them were developed for animal miRNAs which are significantly different from plant miRNAs in the target recognition process (9,10). For example, an animal miRNA generally requires loose complementarity in about first eight nucleotides of the miRNA, while a plant miRNA demands the whole miRNA mature sequence to be near perfectly aligned with its mRNA target. Secondly, an animal miRNA tends to inhibit target gene's expression at the translational level, whereas a plant miRNA directly cleaves its target transcript. In addition, an animal miRNA inclines to recognize 3′-UTR region of its target mRNA and a plant miRNA usually has no preference in terms of position. Recent discoveries suggest that plant miRNA may inhibit gene expression at the translational level (11),

*To whom correspondence should be addressed. Tel: +1 580 224 6725, Fax: +1 580 224 6692; Email: pzhao@noble.org

though it seems to utilize a different recognition pattern compared with a typical animal miRNA's action (see the next section). These differences therefore demand the development of separate plant small RNA target analysis tools.

A regular expression (a computer language that describes matching strings in text) like pattern matching program, PatScan (12), was adopted in identifying the miRNA targets in rice and *Arabidopsis* (13,14). In order to utilize PatScan, users need first prepare regular expression-style pattern sets for an input miRNA sequence, then search target sequences with matching pattern(s) in a candidate mRNA sequence dataset. However, the PatScan program was not designed for the end users such as biologists; extra training and programming skills are usually required in order to successfully install this UNIX-style program and generate comprehensive pattern sets for large-scale systematical miRNA target analysis. Xie *et al.* (15) developed a BLAST-based target search program, miRNAassist. More recently, Fahlgren and Carrington (16) described a pipeline for plant miRNA target prediction using the FASTA program and Perl scripts for matching and scoring. These programs require local installation on a standalone computer and are not designed for high-throughput computing; therefore are not suitable for genome-scale analysis. In addition, the performance of BLAST-like programs are controversial; for example, our study indicated that NCBI BLASTN may miss up to 70% potential targets, because these programs were designed to align long sequences, such as the Expressed Sequence Tags (ESTs) and genomic sequences (8).

Zhang (19) introduced an online analysis tool, miRU, for plant miRNA target analysis. The miRU adopted Smith–Waterman algorithm to search the optimal alignment; it also provides a simple user-friendly web interface and outputs an easily understandable list of matching results, which makes miRU a popular plant miRNA target analysis tool (17,18). However, miRU only accepts one sequence at a time for analysis (lacks high-throughput analysis capability) and the user can only search target candidates in the preloaded libraries. These shortcomings have limited its application in genomic studies, such as analyzing the popular next-generation sequences (19).

The above-mentioned plant miRNA target analysis tools generally focus on the complementarity between small RNA and target transcript. The accessibility of target site on mRNA to a small RNA, determined by secondary structure of mRNA around the target site, has been proved to be an important factor in target recognition (20–23). Incorporating such target-site accessibility evaluation to small RNA target analysis was reported to significantly improve the prediction accuracy (6).

Brodersen *et al.* (11) reported that miRNA translation inhibition might widely spread in plants. However to date, there is no reported plant small RNA target prediction tool that is capable of distinguishing the newly discovered mechanism from the well-accepted post-transcriptional inhibitions.

Here, we present a new plant small RNA target analysis web server, psRNATarget (http://plantgrn.noble.org/psRNATarget/). The psRNATarget integrates two important analysis functions: (i) reverse complementary matching between small RNA and target transcript using a proven scoring schema, and (ii) target-site accessibility evaluation by calculating unpaired energy (UPE) required to 'open' secondary structure around a small RNA target site on the mRNA. This server incorporates recent discoveries in plant small RNA target recognition, e.g. it distinguishes translational inhibition and post-transcriptional inhibition (11), and it reports the number of small RNA/target site pairs that are reportedly associated with small RNA recognition activity to the target transcript. The psRNATarget is designed for high-throughput analysis of next-generation data by implementing a distributed computing pipeline which runs on a Linux cluster at back-end. The server front-end integrates three user-friendly interfaces to accept user-submitted or preloaded small RNAs and transcript sequences and outputs a comprehensive list of small RNAs and matching target sites on candidate transcripts along with built-in online tools for batch downloading, key word searching and the results filtering.

PRINCIPLES AND psRNATarget BACK-END PIPELINE IMPLEMENTATION

Complementarity

psRNATarget evaluates complementarity between small RNA and target gene transcript using the scoring schema originally applied by miRU (19). Instead of using the NCBI BLAST program, we employed a popular Smith–Waterman (24) implementation, ssearch (Version 36.x) (25), in our back-end pipeline since the latter warrants finding the most alignments between very short small RNA sequences and the mRNA sequences.

Multiplicity of target sites

Another good reason for adopting the ssearch program is that it has capability of returning multiple alignments (ssearch versions 36.x and latter releases) for each small RNA/target transcript pair, unlike most of other Smith–Waterman implementations that only return the optimal alignment. Returning multiple optimal alignments enable the reporting of multiple target sites for each small RNA/target transcript pair. This so-called multiplicity of target sites is especially relevant to the biogenesis of siRNA because existence of dual target sites of a miRNA on a specific target transcript has been reported to be an effective trigger of ta-siRNA biogenesis from a TAS (transacting-siRNA) precursor gene (21,26).

Target site accessibility

The psRNATarget server analyzes target accessibility using the RNAup program in Vienna Package (27). RNAup calculates UPE which is the energy required to 'open' secondary structure around target site on mRNA. Less energy represents higher possibility to be an effective target site because the secondary structures may prevent small RNA and target site from contacting. Considering the bulk of RNA-induced silencing complex (RISC), both

5′- and 3′-flanks of target site are included for target accessibility evaluation (22).

Translational inhibition

In plants, it has been observed that mismatches occurred around the center of miRNA/mRNA complementary region tend to disable the cleavage activity of RISC; however, the binding of mRNA to RISC still can block gene expression at the translational level (11). The psRNATarget server reports translational inhibition potential when a mismatch is detected in the central complementary region of the small RNA sequence.

Parallel computing analysis pipeline

Both the Smith–Waterman-based alignment using the search program and the UPE calculation using the RNAup program are very computational intensive. Although these programs produce much more accurate results, they significantly impact the analysis throughput.

We greatly enhanced the analysis throughput by developing an efficient back-end pipeline on the basis of an in-house developed distributed computing platform, namely BioGrid. Upon user submission, the master node of BioGrid system divides the submitted miRNA datasets into multiple subsets and transfers these subsets as well as specified target transcript library to remote computing nodes in a Linux Cluster. Next, the master node remotely calls and monitors analytic programs in these computing nodes, and finally the master node collects outputs when these analysis jobs are completed. The communications between master node and computing nodes are mainly through the Linux SSH (Secure Shell) channel. The psRNATarget back-end system, including the BioGrid platform, was written in Java and Groovy languages.

WEB INTERFACES OF psRNATarget

Input

The server front-end integrates three simplified user-friendly interfaces to accept user-submitted sequences and selection of a preloaded miRNAs and transcript sequences for analysis, i.e. (i) searching user-submitted small RNAs against preloaded transcripts; (ii) searching preloaded small RNAs (miRNAs by species) against user-submitted transcripts; and (iii) searching user-submitted small RNAs against the user-submitted transcripts.

In each input interface, default parameters are suggested and preloaded based on our literature analysis;

however, users may adjust the behavior of back-end pipeline by changing the parameters. Briefly, maximum expectation for complementarity and UPE (maximum energy to unpair the target site) for target accessibility analysis may be decreased to retrieve more stringent prediction results; more potential target sequences might be missed, though (Figure 1, arrows A and C). The current default values of both parameters were suggested based on our benchmark test (see 'Performance' section). The *hspsize* is the length of scoring region for complementary analysis; users are advised to reduce it to the shortest length of small RNAs if the submitted small RNAs are shorter than default *hspsize* value (20 nt). Otherwise, those small RNAs shorter than *hspsize* will be skipped in target analysis (Figure 1, arrow B). The two flanking sizes (lengths of the left and right flanking sequences of the target site) are also adjustable in the target accessibility analysis (Figure 1, arrow D). The users may also adjust the range of central region in which any detected mismatch will be considered as a trigger of translational inhibition (Figure 1, arrow E).

Output

After each successful submission, the end user will be provided a unique URL to trace the analysis progress or check final results at any time. Once the submitted analysis is completed, the psRNATarget server lists the details of the potential small RNA/target site pairs page by page with a comprehensive query and sort tools on the top of each output page for user to easily browse through the results (Figure 2). In addition, psRNATarget allows users to download the entire results in a tab-delimited text file, which is very critical for large-scale data analysis.

PERFORMANCE

The psRNATarget searches target gene based on both complementarity scoring analysis and secondary structure analysis. We demonstrate its performance by predicting target genes of 10 published *Arabidopsis thaliana* miRNAs (Supplementary Data 1) and comparing those predicted target genes to the experimentally validated targets reported in literature. With the default parameters (Expectation ≤ 3, a slightly relaxed threshold), psRNATarget found 92 target candidates (Supplementary Data 2) in TAIR9 cDNA library for the 10 miRNAs, which includes all of 46 validated target gene (100% coverage rate) at 50.0% potential false positive prediction rate. With a more stringent cut-off threshold (Expectation ≤ 2),
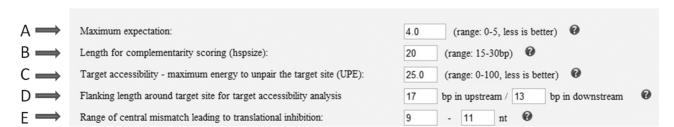

Figure 1. A set of parameters for adjusting the behavior of back-end pipeline of psRNATarget server.

Figure 2. A list of comprehensive miRNA/target site pairs along with query and sort tools on top of each output page.

psRNATarget detected 52 target candidates. Of them, 38 genes have been reportedly validated by 5′-RACE technology, which covers 82.6% of validated target genes at 26.9% potential false positive prediction rate. These results indicate that psRNATarget is able to systematically identify target transcripts; and users may trade their preference on higher prediction coverage or lower false positive prediction rate using different thresholds.

One of the popular applications for psRNATarget is to search target genes in transcript library for small RNAs sequenced by the next-generation technology (28,29). In a performance test, the published small RNA dataset from *Arabidopsis* Small RNA Project (http://asrp.cgrb.oregonstate.edu/), which consists of around 206 000 small RNAs, was submitted to search against the *Arabidopsis* TAIR9 transcript database (http://www.arabidopsis.org/). The psRNATarget took 1 h 54 min to complete the whole analytic procedure running on a Linux cluster equipped with 264 cores (52 AMD opteron processors) and generated around 2.5 million small RNA/target site pairs using default parameter values except that the expectation cutoff value was set ≤4 to generate large number of matching pairs. This benchmark results indicate that psRNATarget is well capable of performing high-throughput analysis for large-scale datasets, such as the next-generation sequencing (NGS) data.

DISCUSSION

Complementarity and target-site accessibility have been proven to be the two key factors in the plant regulatory small RNA (miRNA and ta-siRNA) target recognition mechanism (19–21). Both factors have been incorporated to improve the analysis of miRNA target genes in *Arabidopsis* (6); however, there is no published tool that is able to evaluate these factors for plant regulatory small RNA target analysis for general purpose. The psRNATarget server successfully integrates two well proven approaches (19,27) to evaluate the above two factors; both approaches have been widely applied and well validated by experiments, which warrant the analytic quality of the psRNATarget server.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online

REFERENCES

1. Reinhart,B.J., Weinstein,E.G., Rhoades,M.W., Bartel,B. and Bartel,D.P. (2002) MicroRNAs in plants. *Genes Dev.*, **16**, 1616–1626.
2. Allen,E., Xie,Z., Gustafson,A.M. and Carrington,J.C. (2005) microRNA-directed phasing during trans-acting siRNA biogenesis in plants. *Cell*, **121**, 207–221.
3. Chen,H.-M., Chen,L.-T., Patel,K., Li,Y.-H., Baulcombe,D.C. and Wu,S.-H. (2010) 22-nucleotide RNAs trigger secondary siRNA biogenesis in plants. *Proc. Natl Acad. Sci. USA*, **107**, 15269–15274.
4. Howell,M.D., Fahlgren,N., Chapman,E.J., Cumbie,J.S., Sullivan,C.M., Givan,S.A., Kasschau,K.D. and Carrington,J.C. (2007) Genome-wide analysis of the RNA-DEPENDENT RNA POLYMERASE6/DICER-LIKE4 pathway in Arabidopsis reveals dependency on miRNA- and tasiRNA-directed targeting. *Plant Cell Online*, **19**, 926–942.
5. Bartel,D.P. (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, **116**, 281–297.
6. Li,X. and Zhang,Y.-Z. (2005) Computational detection of microRNAs targeting transcription factor genes in Arabidopsis thaliana. *Comput. Biol. Chem.*, **29**, 360–367.
7. Chaudhuri,K. and Chatterjee,R. (2007) MicroRNA detection and target prediction: integration of computational and experimental approaches. *DNA Cell Biol.*, **26**, 321–337.

8. Dai,X., Zhuang,Z. and Zhao,P.X. (2010) Computational analysis of miRNA targets in plants: current status and challenges. *Brief. Bioinformatics*, **12**, 115–121.

9. Li,L., Xu,J., Yang,D., Tan,X. and Wang,H. (2010) Computational approaches for microRNA studies: a review. *Mamm. Genome*, **21**, 12.

10. Lim,L.P., Lau,N.C., Weinstein,E.G., Abdelhakim,A., Yekta,S., Rhoades,M.W., Burge,C.B. and Bartel,D.P. (2003) The microRNAs of Caenorhabditis elegans. *Genes Dev.*, **17**, 991–1008.

11. Brodersen,P., Sakvarelidze-Achard,L., Bruun-Rasmussen,M., Dunoyer,P., Yamamoto,Y.Y., Sieburth,L. and Voinnet,O. (2008) Widespread translational inhibition by plant miRNAs and siRNAs. *Science*, **320**, 1185–1190.

12. Dsouza,M., Larsen,N. and Overbeek,R. (1997) Searching for patterns in genomic data. *Trends Genet.*, **13**, 497–498.

13. Jones-Rhoades,M.W. and Bartel,D.P. (2004) Computational identification of plant microRNAs and their targets, including a stress-induced miRNA. *Mol. Cell*, **14**, 787–799.

14. Rhoades,M.W., Reinhart,B.J., Lim,L.P., Burge,C.B., Bartel,B. and Bartel,D.P. (2002) Prediction of plant microRNA targets. *Cell*, **110**, 513–520.

15. Xie,F.L., Huang,S.Q., Guo,K., Xiang,A.L., Zhu,Y.Y., Nie,L. and Yang,Z.M. (2007) Computational identification of novel microRNAs and targets in Brassica napus. *FEBS Lett.*, **581**, 1464–1474.

16. Fahlgren,N. and Carrington,J.C. (2010) miRNA target prediction in plants. *Methods Mol. Biol.*, **592**, 51–57.

17. Kawashima,C.G., Yoshimoto,N., Maruyama-Nakashita,A., Tsuchiya,Y.N., Saito,K., Takahashi,H. and Dalmay,T. (2009) Sulphur starvation induces the expression of microRNA-395 and one of its target genes but in different cell types. *Plant J.*, **57**, 313–321.

18. Meng,Y., Shao,C. and Chen,M. (2011) Toward microRNA-mediated gene regulatory networks in plants. *Brief. Bioinformatics*, doi:10.1093/bib/bbq091.

19. Zhang,Y. (2005) miRU: an automated plant miRNA target prediction server. *Nucleic Acids Res.*, **33**, W701–W704.

20. Hausser,J., Landthaler,M., Jaskiewicz,L., Gaidatzis,D. and Zavolan,M. (2009) Relative contribution of sequence and structure features to the mRNA binding of Argonaute/EIF2C-miRNA complexes and the degradation of miRNA targets. *Genome Res.*, **19**, 2009–2020.

21. Brodersen,P. and Voinnet,O. (2009) Revisiting the principles of microRNA target recognition and mode of action. *Nat. Rev. Mol. Cell Biol.*, **10**, 141–148.

22. Kertesz,M., Iovino,N., Unnerstall,U., Gaul,U. and Segal,E. (2007) The role of site accessibility in microRNA target recognition. *Nat. Genet.*, **39**, 1278–1284.

23. Marin,R.M. and Vanicek,J. (2011) Efficient use of accessibility in microRNA target prediction. *Nucleic Acids Res.*, **39**, 19–29.

24. Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197.

25. Pearson,W.R. (1991) Searching protein sequence libraries: comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms. *Genomics*, **11**, 635–650.

26. Axtell,M.J., Jan,C., Rajagopalan,R. and Bartel,D.P. (2006) A two-hit trigger for siRNA biogenesis in plants. *Cell*, **127**, 565–577.

27. Muckstein,U., Tafer,H., Hackermuller,J., Bernhart,S.H., Stadler,P.F. and Hofacker,I.L. (2006) Thermodynamics of RNA-RNA binding. *Bioinformatics*, **22**, 1177–1182.

28. Ruan,M.-B., Zhao,Y.-T., Meng,Z.-H., Wang,X.-J. and Yang,W.-C. (2009) Conserved miRNA analysis in Gossypium hirsutum through small RNA sequencing. *Genomics*, **94**, 263–268.

29. Schreiber,A., Shi,B.-J., Huang,C.-Y., Langridge,P. and Baumann,U. (2011) Discovery of barley miRNAs through deep sequencing of short reads. *BMC Genomics*, **12**, 129.