Nucleic Acids Research, 2005, Vol. 33, Web Server issue ***W701–W704***
doi:10.1093/nar/gki383

miRU: an automated plant miRNA target prediction server

Yuanji Zhang*

Plant Biology Division, The Samuel Roberts Noble Foundation, Ardmore, OK 73402, USA

Received February 3, 2005; Revised and Accepted March 9, 2005

ABSTRACT

MicroRNAs (miRNAs) play important roles in gene expression regulation in animals and plants. Since plant miRNAs recognize their target mRNAs by near-perfect base pairing, computational sequence similarity search can be used to identify potential targets. A web-based integrated computing system, miRU, has been developed for plant miRNA target gene prediction in any plant, if a large number of sequences are available. Given a mature miRNA sequence from a plant species, the system thoroughly searches for potential complementary target sites with mismatches tolerable in miRNA–target recognition. True or false positives are estimated based on the number and type of mismatches in the target site, and on the evolutionary conservation of target complementarity in another genome which can be selected according to miRNA conservation. The output for predicted targets, ordered by mismatch scores, includes complementary sequences with mismatches highlighted in colors, original gene sequences and associated functional annotations. The miRU web server is available at http://bioinfo3.noble.org/miRU.htm.

INTRODUCTION

MicroRNAs (miRNAs) are endogenously encoded small RNAs that can regulate gene expressions by base-pairing to protein-coding mRNAs for degradation or translation repression. Numerous miRNAs have been identified from genomes of many animals and plants, such as fruit fly, nematode, zebrafish, chicken, mouse, human, Arabidopsis, rice and maize. miRNA genes are abundant in humans, estimated to account for ~1% of the total predicted genes. In Arabidopsis, at least 43 distinct miRNA families consisting of 111 members have been reported and archived in 'The miRNA Registry' thus far (http://www.sanger.ac.uk/Software/Rfam/mirna/index.shtml) (1). Although the function of most miRNAs remains unknown, a number of miRNAs have been shown to play important roles in developmental timing, cell death, cell proliferation, hematopoiesis and patterning of the nervous system in animals, and stress responses, and leaf and flower development in plants (2–6).

Finding regulatory mRNA targets is essential to understanding the biological functions of miRNAs. Different methods are needed to predict animal and plant miRNA targets. While miRNA–target duplex free energy may be important for animal miRNA target prediction (7,8), plant miRNA targets can be predicted by sequence similarity since plant miRNA seems to bind almost perfectly to its cognate mRNA (7,9). Computational tools have been developed to predict plant miRNA targets (9–11), but none is in the web server format. Rhoades *et al*. (9) used PatScan (12) to predict plant miRNA targets with ≤3 mismatches. Jones-Rhoades and Bartel (11) used their own unpublished programs, together with PatScan, and the prediction seems to be more comprehensive. Wang *et al*. (10) deployed Smith–Waterman algorithm in miRNA target prediction, but failed to detect all previously identified targets (10). Since most biology laboratories involved in plant miRNA research may not have necessary bioinformatic resources for target prediction, a publicly accessible web application for plant miRNA target prediction has been developed. The tool allows systematic search for miRNA complementary targets in any plant whose genome sequence or a large number of expressed sequence tags (ESTs) are available. Backed by an exhaustive search algorithm, the tool is able to find all potential targets with the given mismatches. False positives are reduced by limiting the number of mismatches and by ensuring the target complementarity conservation in another plant species (11).

INPUT TO THE SERVER

The server has a user-friendly and intuitive input interface, as shown in Figure 1. The user is required to enter a mature miRNA sequence in 5′→3′direction. Although miRNAs are usually 21–24 nt (4), the input sequence can be in the range of 19–28 nt in length to accommodate an siRNA input, as the tool can also be used to search for siRNA targets and off-targets.

*Tel: +1 580 224 6726; Fax: +1 580 224 6692; Email: yjzhang@noble.org

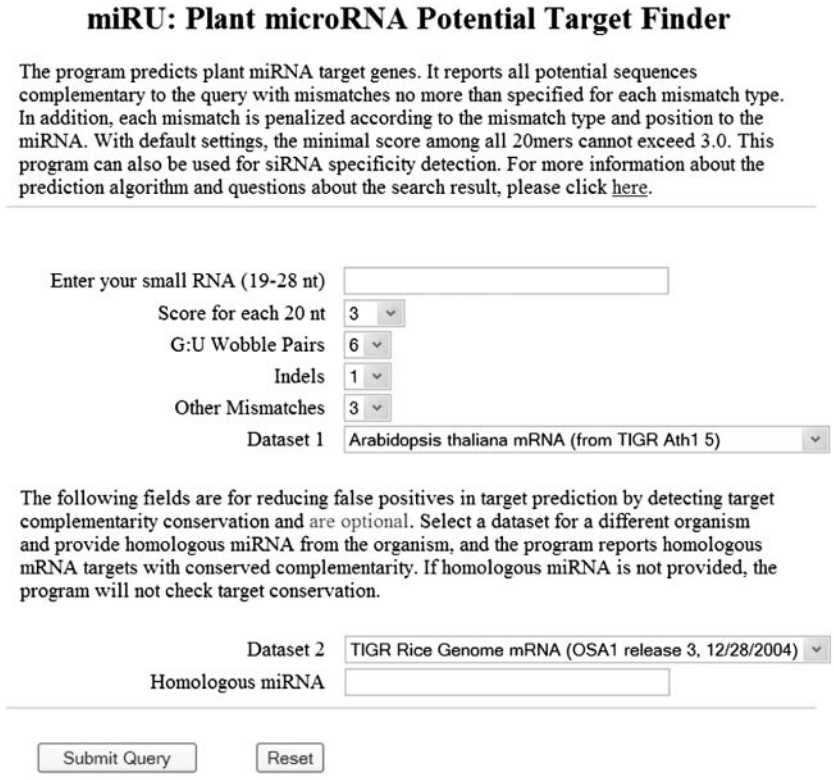


Figure 1. Data input web interface for plant miRNA target prediction.

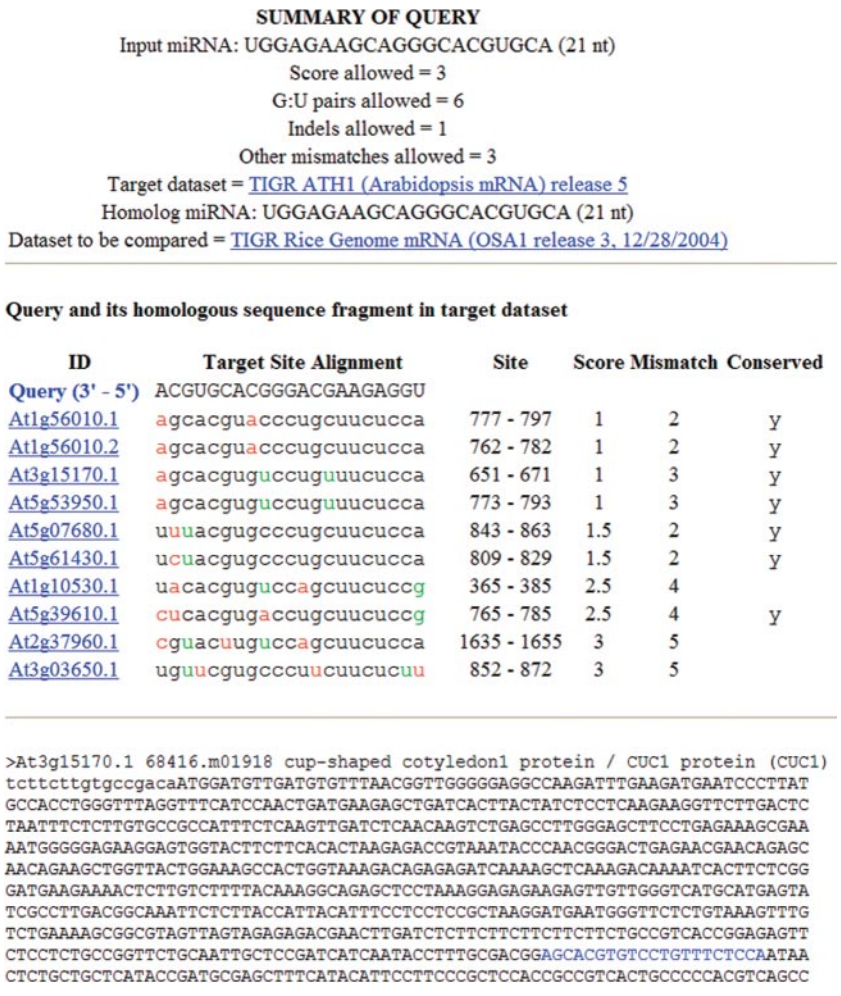
To predict target genes, the user has to specify an mRNA dataset for the intended organism. Currently, the system includes genome mRNAs or ESTs and other transcripts-assembled Gene Indices (13) for 28 plant species downloaded from The Institute for Genome Research (TIGR) at http://www.tigr.org/. With the above input information, a Perl script at the backend will then do an exhaustive sequence similarity search, using an algorithm modified from BLAST (14) (see Additional File 1).

To reduce false positives in predicted targets, the user can limit the number of mismatches, which are classified into three types and are assigned different scores; the higher scores are for more detrimental mismatches for miRNA function: G:U wobble pairings (each assigned 0.5 scores), insertions/deletions (indels) (2.0) and all other non-canonical Watson–Crick pairings (1.0). The total score for an alignment is calculated based on 20 nt. When the query is longer than 20 nt, scores for all possible consecutive 20 nt subsequences are computed and the minimum score is output as the total score for the query-subject alignment. Since target complementarity to the miRNA 5′ end seems to be critical to the target site function (15–18), any mismatch other than G:U wobble in positions 2–7 at the 5′ end is further penalized 0.5 points in the score.

Based on the observation that both miRNAs and their target sites are evolutionarily conserved across genomes (18–20), the conservation of target complementarity in another genome can be used to further reduce false positives in plant miRNA target prediction (11). Furthermore, such analysis will also provide useful information about conserved regulatory roles of homologous miRNAs in different species. To use this strategy in the server, the information of the homologous miRNA and the mRNA dataset of the second genome should be provided for the system to do another search. Then the system compares potential targets to find whether homologous genes are predicted to be targeted by the homologous miRNAs in both genomes. Genes are considered to be homologous if they share $\geqslant$1 Pfam domains (21). All mRNA datasets are preprocessed by aligning to Pfam-A seed domain sequences (Pfam 16.0, which contains 7677 families, available at http://www.sanger.ac.uk/Software/Pfam/). For Arabidopsis and rice genome mRNA datasets, the corresponding protein datasets are used for functional domain identification using HMMER (22) with *E*-value $\leqslant$0.1 as the significance level. Since HMMER does not allow DNA–protein comparisons, all gene index datasets are searched against Pfam-A seed domain sequences using blastx program (14) with *E*-value cut-off of 10^{-5}. TIGR's 'Eukaryotic Gene Orthologs' dataset (23) is also used for determining homology relationships in the Gene Index datasets. The search results are parsed and stored in a MySQL database to facilitate the comparisons of target conservation in any two genomes.

OUTPUT TO THE USER

The output report consists of three parts (Figure 2). The first part is a summary of search input parameters, including the query sequence, mismatches allowed and target dataset. The next section is a list of predicted complementary targets displayed in the order of mismatch scores. Information shown for each predicted target includes gene identifier, target site position, mismatch score, number of mismatches and target



SUMMARY OF QUERY

Input miRNA: UGGAGAAGCAGGGCACGUGCA (21 nt)
Score allowed = 3
G:U pairs allowed = 6
Indels allowed = 1
Other mismatches allowed = 3
Target dataset = TIGR ATH1 (Arabidopsis mRNA) release 5
Homolog miRNA: UGGAGAAGCAGGGCACGUGCA (21 nt)
Dataset to be compared = TIGR Rice Genome mRNA (OSA1 release 3, 12/28/2004)

Query and its homologous sequence fragment in target dataset

ID	Target Site Alignment	Site	Score	Mismatch	Conserved
Query (3' - 5')	ACGUGCACGGGACGAAGAGGU				
At1g56010.1	agcacguacccugcuucucca	777 - 797	1	2	y
At1g56010.2	agcacguacccugcuucucca	762 - 782	1	2	y
At3g15170.1	agcacguguccuguuucucca	651 - 671	1	3	y
At5g53950.1	agcacguguccuguuucucca	773 - 793	1	3	y
At5g07680.1	uuuacgugcccugcuucucca	843 - 863	1.5	2	y
At5g61430.1	ucuacgugcccugcuucucca	809 - 829	1.5	2	y
At1g10530.1	uacacguguccagcuucuccg	365 - 385	2.5	4	
At5g39610.1	cucacgugaccugcuucuccg	765 - 785	2.5	4	y
At2g37960.1	cguacuuguccagcuucucca	1635 - 1655	3	5	
At3g03650.1	uguucgugcccuucuucucuu	852 - 872	3	5	

>At3g15170.1 68416.m01918 cup-shaped cotyledon1 protein / CUC1 protein (CUC1)
tcttcttgtgccgacaATGGATGTTGATGTGTTTAACGGTTGGGGGAGGCCAAGATTTGAAGATGAATCCCTTAT
GCCACCTGGGTTTAGGTTTCATCCAACTGATGAAGAGCTGATCACTTACTATCTCCTCAAGAAGGTTCTTGACTC
TAATTTCTCTTGTGCCGCCATTTCTCAAGTTGATCTCAACAAGTCTGAGCCTTGGGAGCTTCCTGAGAAAGCGAA
AATGGGGGAGAAGGAGTGGTACTTCTTCACACTAAGAGACCGTAAATACCCAACGGGACTGAGAACGAACAGAGC
AACAGAAGCTGGTTACTGGAAAGCCACTGGTAAAGACAGAGAGATCAAAAGCTCAAAGACAAAATCACTTCTCGG
GATGAAGAAAACTCTTGTCTTTTACAAAGGCAGAGCTCCTAAAGGAGAGAAGAGTTGTTGGGTCATGCATGAGTA
TCGCCTTGACGGCAAATTCTCTTACCATTACATTTCCTCCTCCGCTAAGGATGAATGGGTTCTCTGTAAAGTTTG
TCTGAAAAGCGGCGTAGTTAGTAGAGAGACGAACTTGATCTCTTCTTCTTCTTCTGCCGTCACCGGAGAGTT
CTCCTCTGCCGGTTCTGCAATTGCTCCGATCATCAATACCTTTGCGACGGAGCACGTGTCCTGTTTCTCCAATAA
CTCTGCTGCTCATACCGATGCGAGCTTTCATACATTCCTTCCCGCTCCACCGCCGTCACTGCCCCCACGTCAGCC

Figure 2. Result page of plant miRNA target prediction.

complementary sequence with mismatches highlighted in colors (green for G·U mismatches, purple for indels and red for all other mismatches). The target is indicated if its complementarity is conserved in another genome. Therefore, the target list includes conserved targets that are highly likely to be true targets. It also includes targets whose counterparts in the second genome are not found. Some of these targets may still be true targets since the dataset to be compared for most plants are ESTs sampled from the genomes and may miss the conserved targets. The last part of the output is the target gene sequences in FASTA format, which includes the definition line for the original functional annotation. The target site in the gene can easily be located as it is highlighted in colors (Figure 2).

To verify the tool, its prediction was compared with two published prediction results (6,11). The prediction of Arabidopsis miRNA targets by Jones-Rhoades and Bartel (11) seems to be highly reliable since more than half of the predicted targets were experimentally verified as true targets. In this work, Arabidopsis miRNAs conserved in rice, as listed in Supplementary Table S1 in Jones-Rhoades and Bartel (11), were used as queries for the tool to predict target genes and the result can be found in Additional File 2. All the reported potential target genes were successfully detected by this tool. Recently, Sunkar and Zhu (6) identified stress-regulated miRNAs from Arabidopsis. They also predicted the potential target genes for these miRNAs using the criteria modified from Rhoades *et al*. (9). The new algorithm detected all their predicted targets. Moreover, the result indicates that Sunkar and Zhu's prediction seems to be incomplete. For example, a total of 23 targets were predicted by Sunkar and Zhu for ten miRNAs identified in their experiment, while this server predicts 203 potential targets in total (see Additional File 3).

SUMMARY

The server aims at predicting plant miRNA targets with the highest sensitivity and selectivity by using a search algorithm which guarantees finding all homologous sequences within given mismatches, and by applying current knowledge about miRNA targets to minimize false positives. As a practical tool, it should aid biologists in plant miRNA research.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENTS

The author would like to acknowledge Drs Richard A. Dixon and Patrick Zhao for critical reading of the manuscript. Financial support for this project was provided by the Samuel Roberts Noble Foundation. Funding to pay the Open Access publication charges for this article was also provided by the Samuel Roberts Noble Foundation.

Conflict of interest statement. None declared.

REFERENCES

1. Griffiths-Jones,S. (2004) The microRNA Registry. *Nucleic Acids Res*., **32**, D109–D111.
2. Bartel,D.P. (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, **116**, 281–297.
3. Carrington,J.C. and Ambros,V. (2003) Role of microRNAs in plant and animal development. *Science*, **301**, 336–338.
4. Dugas,D.V. and Bartel,B. (2004) MicroRNA regulation of gene expression in plants. *Curr. Opin. Plant Biol*., **7**, 512–520.
5. Ambros,V. (2004) The functions of animal microRNAs. *Nature*, **431**, 350–355.
6. Sunkar,R. and Zhu,J.-K. (2004) Novel and stress-regulated microRNAs and other small RNAs from Arabidopsis. *Plant Cell*, **16**, 2001–2019.
7. Lai,E.C. (2004) Predicting and validating microRNA targets. *Genome Biol*., **5**, 115.
8. Rehmsmeier,M., Steffen,P., Hochsmann,M. and Giegerich,R. (2004) Fast and effective prediction of microRNA/target duplexes. *RNA*, **10**, 1507–1517.
9. 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.
10. Wang,X.J., Reyes,J.L., Chua,N.H. and Gaasterland,T. (2004) Prediction and identification of *Arabidopsis thaliana* microRNAs and their mRNA targets. *Genome Biol*., **5**, R65.
11. 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.
12. Dsouza,M., Larsen,N. and Overbeek,R. (1997) Searching for patterns in genomic data. *Trends Genet*., **13**, 497–498.
13. Quackenbush,J., Liang,F., Holt,I., Pertea,G. and Upton,J. (2000) The TIGR Gene Indices: reconstruction and representation of expressed gene sequences. *Nucleic Acids Res*., **28**, 141–145.
14. Altschul,S., Thomas,L., Alejandro,A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*., **25**, 3389–3402.
15. Mallory,A.C., Reinhart,B.J., Jones-Rhoades,M.W., Tang,G., Zamore,P.D., Barton,M.K. and Bartel,D.P. (2004) MicroRNA control of PHABULOSA in leaf development: importance of pairing to the microRNA 5′ region. *EMBO J*., **23**, 3356–3364.
16. Brennecke,J., Stark,A., Russell,R.B. and Cohen,S.M. (2005) Principles of microRNA–target recognition. *PLoS Biol*., **3**, e85.
17. Lim,L.P., Lau,N.C., Garrett-Engele,P., Grimson,A., Schelter,J.M., Castle,J., Bartel,D.P., Linsley,P.S. and Johnson,J.M. (2005) Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature*, **433**, 769–773.
18. Stark,A., Brennecke,J., Russell,R.B. and Cohen,S.M. (2003) Identification of Drosophila microRNA targets. *PLoS Biol*., **1**, E60.
19. Juarez,M.T., Kui,J.S., Thomas,J., Heller,B.A. and Timmermans,M.C.P. (2004) microRNA-mediated repression of rolled leaf1 specifies maize leaf polarity. *Nature*, **428**, 84–88.
20. Bonnet,E., Wuyts,J., Rouze,P. and Van de Peer,Y. (2004) Detection of 91 potential conserved plant microRNAs in *Arabidopsis thaliana* and *Oryza sativa* identifies important target genes. *Proc. Natl Acad. Sci. USA*, **101**, 11511–11516.
21. Bateman,A., Coin,L., Durbin,R., Finn,R.D., Hollich,V., Griffiths-Jones,S., Khanna,A., Marshall,M., Moxon,S., Sonnhammer,E.L. *et al*. (2004) The Pfam protein families database. *Nucleic Acids Res*., D138–D141.
22. Durbin,R., Eddy,S., Krogh,A. and Mitchison,G. (1998) *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, Cambridge, UK.
23. Lee,Y., Sultana,R., Pertea,G., Cho,J., Karamycheva,S., Tsai,J., Parvizi,B., Cheung,F., Antonescu,V. and White,J. (2002) Cross-referencing eukaryotic genomes. *Genome Res*., **12**, 493–502.