Bioinformatics Advance Access published April 28, 2010

TAPIR, a web server for the prediction of plant microRNA targets, including target mimics

Eric Bonnet [1,2], Ying He [1,2], Kenny Billiau [1,2] and Yves Van de Peer [1,2*]

[1]Department of Plant Systems Biology, VIB, Technologiepark 927, B-9052 Gent, Belgium.
[2]Department of Molecular Genetics, Ghent University, Technologiepark 927, B-9052 Gent, Belgium.

Associate Editor: Prof. Ivo Hofacker

ABSTRACT

Summary: We present a new web server called TAPIR, designed for the prediction of plant microRNA targets. The server offers the possibility to search for plant miRNA targets using a fast and a precise algorithm. The precise option is much slower but guarantees to find less perfectly paired miRNA - target duplexes. Furthermore, the precise option allows the prediction of target mimics, which are characterized by a miRNA - target duplex having a large loop, making them undetectable by traditional tools.

Availability: The TAPIR web server can be accessed at: http://bioinformatics.psb.ugent.be/webtools/tapir

Contact: yves.vandepeer@psb.vib-ugent.be

INTRODUCTION

MicroRNAs (miRNAs) constitute a prominent class of small non-coding RNAs that regulate gene expression at the post-transcriptional level. In plants, miRNAs are key regulators involved in various developmental, stress and cellular responses (Voinnet, 2009). With the rise of next-generation sequencing data, new miRNAs are uncovered in various plant genomes at a rapid pace. One of the major challenges is then to determine their function, and a crucial step towards this goal is the identification of the miRNAs targets. Candidate targets can be identified through base pair complementarity of the miRNA sequence to mRNA sequences (Rhoades *et al*., 2002). The miRNA:mRNA duplexes were originally thought to have few mismatches, G:U pairs and bulges, but there is now increasing evidence that less perfect hybridizations might also be functional, leading in some cases to translational repression rather than cleavage of the target mRNA (Brodersen *et al*., 2008; Dugas and Bartel, 2008; Brodersen and Voinnet, 2009). Recent reports have also revealed the existence in plants of a phenomenon called miRNA target mimicry, where miRNA:mRNA duplexes with a large bulge around the cleavage site actually sequester miRNAs, resulting in an inhibition of the miRNA activity (Franco-Zorilla *et al*., 2007). We have designed and implemented TAPIR, a novel web server dedicated to the prediction of plant miRNAs, including miRNA target mimics.

*to whom correspondence should be addressed

METHODS

For the detection of miRNA:mRNA duplexes, we use two different previously published algorithms. The first is the classical FASTA local alignment program, which is very fast but cannot detect the duplexes having a lot of bulges and/or mismatches (Pearson, 2004). The second algorithm is RNAhybrid, an algorithm for a precise detection of the miRNA:mRNA duplexes (Rehmsmeier *et al*., 2004).

FASTA search engine

For the FASTA program, the sequences are reverse complemented and submitted as queries, while the mRNA sequences are submitted as targets. In order to maximize the sensitivity, the program is launched with a k-tuple subword size of 1 and an e-value cutoff set to 150. The results of the FASTA search are then parsed to calculate the score of each duplex. Previous work has shown that the miRNA:mRNA duplex free energy ratio is also an important parameter to define valid plant miRNA targets. This parameter is defined by the ratio of the free energy of the duplex to the free energy of the same duplex having only perfect matches (Allen *et al*., 2005; Schwab *et al*., 2005). In order to calculate this parameter, the miRNA and the mRNA sequences of the duplexes are linked by a short sequence forming a loop structure. The free energy of the structure is then calculated using the ViennaRNA package (Hofacker *et al*., 1994).

RNAhybrid search engine

RNAhybrid is an extension of the classical RNA secondary structure prediction algorithm (Rehmsmeier *et al*., 2004). Using a dynamic programming algorithm, the program calculates the minimum free energy hybridizations of all possible start positions of the miRNA within the target sequence, with a restriction on the length of internal bulges and loops. Furthermore, the free energy of the duplex is precisely calculated and is not sensitive to artifacts of miRNA-mRNA concatenation (Rehmsmeier *et al*., 2004). RNAhybrid has been successfully used to predict novel plant miRNAs (Alves *et al*., 2009). The disadvantage of using this algorithm is that it is considerably slower compared to heuristic local alignment programs like FASTA. The program is used with default values, except that we request 10 hits per target (option -b 10) and limit the size of gaps to 5 nt on each side of the duplex (option -u 5). The search results are further parsed to calculate the target score or to look for miRNA target mimicry pattern.

miRNA target score

The score calculated for each miRNA:mRNA duplex is the same as the one used by Allen and colleagues (Allen *et al*., 2005), which was derived from previous studies (Rhoades *et al*., 2002; Schwab *et al*., 2005). The score is taking into account the number of mismatches, the number of gaps (introduced by bulges and loop structures) and the number of G:U pairs. Moreover, several studies have shown the importance of a ”seed” or ”core” region of the duplex, in the 5’ region of the miRNA. This region is

significantly depleted in mismatches and bulges for valid miRNA targets. The score is taking this fact into account by applying a penalization factor to mismatches, gaps and G:U pairs located between positions 2 and 12 of the miRNA sequence.

$$S = Nm + Ng + (0.5 \cdot Nu) + (2 \cdot Nsm) + (2 \cdot Nsg) + Nsu \quad (1)$$

where Nm, Ng and Nu are the number of mismatches, gaps and G:U pairs outside the "seed" region and Nsm, Nsg and Nsu are the same parameters within the "seed" region.

miRNA target mimicry search

MiRNA target mimicry miRNA:mRNA duplexes are characterized by the presence of a large bulge located within the typical cleavage site for plant miRNAs, i.e. between the nucleotides 10 and 11 of the miRNA sequence (Franco-Zorilla *et al*., 2007). Of course, such patterns can not be detected by local alignment programs like FASTA, but the sensitivity of the RNAhybrid program allow to detect such weak hybridizations. We used the duplex between miR399 and its target mimic *IPS1* (Franco-Zorilla *et al*., 2007) as a template to define miRNA target mimic parameters. We retain duplexes having a bulge of 3 nt between the nucleotides 10 and 11 of the miRNA, for which the nucleotides 10 and 11 are engaged in a Watson-Crick base pair or a G:U pair and for which the free energy ratio of the duplex is higher than a certain cutoff value (by default 0.7).

RESULTS AND DISCUSSION

To test and compare the TAPIR web server algorithms to existing web tools , we have compiled a reference set of 102 miRNA - target pairs that have been validated experimentally (Allen *et al*., 2005; Addo-Quaye *at al*., 2008; German *et al*., 2008). The sensitivity is defined as the number of pairs that each program can recover from the reference set. The specificity is defined by the fraction of predicted pairs that do not belong to the reference set to the total number of pairs predicted, assuming that the reference set covers all the true miRNA - target pairs. Of course, it's very likely that current efforts to validate predicted targets may have missed a true positive, even for high-throughput approaches. A true target might be missed because the choice of the experimental conditions does not match the conditions in which the miRNA or the target is active, like for example specific biotic or abiotic stresses. Plant miRNAs have indeed been shown to be involved in several stress responses (Sunkar and Zhu, 2004). All the validation methods for plant miRNAs so far are looking for traces of cleavage products, so one can not rule out that the target might be repressed at the translational level rather than cleaved. Recent studies have shown that translational repression mediated by miRNAs is more frequent than previously thought in plants (Brodersen *et al*., 2008; Dugas and Bartel, 2008; Brodersen and Voinnet, 2009). As a consequence, the specificity values calculated in this study might be biased towards high numbers, because some false negatives might in fact be true positives. However, by looking at the difference between two values rather than the absolute values themselves, we can estimate the fraction of false positives between two different methods or between different parameters values for the same method. We performed the sensitivity and specificity tests for the TAPIR fast and precise methods using various parameter values, and we also included in this test two other web tools designed for the prediction of plant miRNA targets, miRU (Zhang, 2005) and Srna target (Moxon *et al*., 2008). All the results from those tests are compiled in the table 1.

The Srna interface does not allow the user to modify any search parameter, so we just used the default parameters. For miRU, the expectation score can be adjusted between different discrete values and we ran the program with the values 3, 4 and 5 (the latter being the maximum value allowed). We ran the TAPIR web tool with the score cutoff values of 4, 5 and 7 and free energy ratio cutoff of 0.7, 0.7 and 0.6 respectively for both the fast (FASTA) and the precise (RNAhybrid) methods. As expected, the fraction of true positive as well as the false positives is increasing when the cutoff parameters are lowered for miRU and both TAPIR methods. The default (and fixed) parameters of the Srna target tool are performing better than the miRU search (score cutoff 3, 4 and 5), having a higher percentage of true positives (82.4 *vs* 75.5) and a lower percentage of false positives (68.3 *vs* 71.6). The TAPIR with the FASTA search engine (score cutoff 4) has a higher rate of true positive (89.2) while keeping the false positives (70.0) to values that are similar to those of Srna (68.3) and miRU (71.6). The TAPIR search with the RNAhybrid engine (score cutoff 4) has the same value of true positives as with the FASTA engine (89.2) but the rate of false positives is higher (78.3), due likely to its increased capability to detect weaker miRNA:mRNA duplexes. If we compare the TAPIR FASTA and RNAhybrid search engines, we can see that the latter is giving equal or higher values for true positives for all score cutoff considered (although the absolute value increase is not dramatic), but this higher accuracy is coming at the cost of higher false positive rates. The RNAhybrid search engine is eventually capable of recovering 100% of the reference set, but with an extremely high number of false positives (data not shown).

We also benchmarked the TAPIR FASTA and RNAhybrid search engines on a set of 5,000 miRNA - target pairs to determine the average time for one miRNA - target comparison. The average time for the FASTA search is $6 \cdot 10^{-4}$ sec, while the RNAhybrid search is 0.144 sec. This makes the FASTA search about 240 times faster.

We have seen that the sensitivity of the RNAhybrid search over the FASTA is only slightly better, with a significant increase in the proportion of false positives. Furthermore we have also seen that the RNAhybrid search is considerably slower, we can thus conclude that the FASTA search engine is better suited for rapid, genome wide searches of plant miRNA targets, while the RNAhybrid search

Table 1. Sensitivity and specificity for plant miRNA prediction programs, tested on a reference set of 102 validated miRNA - target pairs.

	Total	Pos.	P	FP
Srna target	265	84	82.4	68.3
miRU exp. sc. ≤ 3	271	77	75.5	71.6
miRU exp. sc. ≤ 4	620	81	79.4	86.9
miRU exp. sc. ≤ 5	922	81	79.4	91.2
Tapir fasta sc. ≤ 4 & rat. ≥ 0.7	303	91	89.2	70.0
Tapir fasta sc. ≤ 5 & rat. ≥ 0.7	488	93	91.2	80.9
Tapir fasta sc. ≤ 7 & rat. ≥ 0.6	1485	97	95.1	93.5
Tapir hybrid sc. ≤ 4 & rat. ≥ 0.7	414	91	89.2	78.3
Tapir hybrid sc. ≤ 5 & rat. ≥ 0.7	926	94	92.2	89.8
Tapir hybrid sc. ≤ 7 & rat. ≥ 0.6	15036	100	98.0	99.3

Abbreviations: exp. sc. = expectation score, sc. = score, rat. = duplex free energy ratio., Total = total number of predicted targets, Pos. = number of reference miRNA - target pairs identified, P = true positive percentage, FP = false positive percentage

might be useful for extensive searches for a particular miRNA sequence, when an increased sensitivity is desirable. Of course, the added value of the RNAhybrid precision is evident with its capacity to predict the weak interactions corresponding to target mimics (Franco-Zorilla *et al*., 2007), prediction that is absolutely impossible with the FASTA search engine.

The TAPIR web server is implemented as a user friendly web interface (no login or email is required), with one tab for the FASTA (”Fast”) and one for the RNAhybrid (”Precise”) search engines. Users can paste or upload fasta files corresponding to the miRNA and target sequences. Users can also set two parameters for the search, the score cutoff value (default 4) and the free energy ratio cutoff value (default 0.7).

By comparison, the Srna (http://srna-tools.cmp.uea.ac.uk/plant/cgi-bin/srna-tools.cgi?rm=input_form&tool=target) interface does not allow any parameter to be changed and only allow to upload the miRNA sequences. The target database has to be chosen from a pre-defined list of more than 15 plant species. Furthermore, the interface does not allow an ”anonymous” search and requires the user to enter a valid e-mail. The search will be started only when the user will click on the link send to his e-mail address. With the miRU enhanced interface (http://bioinfo3.noble.org/psRNATarget/), the user can leave his email as an option and can choose to either upload both miRNA and target sequences, upload the miRNA sequences and search a set of pre-defined target sequences (more than 15 plant genomes) or even upload the target sequences and use a set of pre-defined plant miRNA sequences. The user can also set two parameters, the expectation score (threshold corresponding to the scoring scheme applied to the miRNA:mRNA duplex) and the maximum circles (iterations of the underlying Smith-Waterman algorithm to search for multiple hits on the same target sequence). The miRU results do not report the free energy value or ratio of the miRNA:mRNA duplexes.

For the TAPIR web interface, once the search is submitted, the user is redirected to a temporary web page indicating the duration of the search until it is completed. The page can be bookmarked by the user for a later check or retrieval of the result. All results are kept for 24 hours on the server. All the miRNA - target pairs for which the score is less than or equal to the cutoff value and the free energy ratio is greater than or equal to the cutoff value are displayed in the results page. Several parameter values are indicated (names, free energy ratio, start position of the duplex on the mRNA, seed and non-seed mismatches, gaps and G:U pairs), plus a full representation of the miRNA:mRNA duplex with an alignment string (Fig. 1A).

A

```
miRNA          ath-miR163
target         AT1G66720.1
score          2
mfe_ratio      0.91
start          333
seed_gap       0
seed_mismatch  0
seed_gu        1
gap            1
mismatch       0
gu             0
miRNA_seq   3' UAGCUUCAAGGUUCAGGAGAAGUU
aln            ||||.|||||||o||||||||||
target_seq 5' AUCG-AGUUCCAGGUCCUCUUCAA
```

B

```
miRNA     ath-miR399a
target    IPS1-AT3G09922.1
mfe       -31.9
mfe ratio 0.7800
start     237
gap       3
mismatch  2
GU        1
miRNA  3' UCCCGUUUAG---AGGAAACCGU
aln       .||||||.|o...||||||||||
target 5' UGGGCAACUUCUAUCCUUUGGCA
```

Fig. 1. Search results examples from the TAPIR web server. (A) miRNA target search. (B) miRNA target mimic search.

With the ”Precise” search comes the ability to look for target mimics. The user only has to check the corresponding box and set the free energy ratio cutoff to proceed. The results are displayed in the same way as for target search (Fig. 1B). The RNAhybrid algorithm being quite slow, we have pre-computed searches for ten different plant genomes and their corresponding miRNAs (the complete list of genomes is available on the website, within the ”manual” pages). The plant miRNAs were downloaded from miRBase (Griffiths-Jones *et al*., 2008). As plant miRNAs can sometimes bind in UTR regions, we have used cDNA sequences when available. The users can use pre-computed data by simply selecting the organism and the miRNA they are interested in. The raw data is then parsed according to the parameters selected by the user.

The TAPIR web server is bringing new features compared to existing solutions (Zhang, 2005; Moxon *et al*., 2008). The ability to use two different search engines, the rich output results featuring a precise calculation of the free energy and free energy ratio, the possibility to look for target mimics and to use pre-computed results should make TAPIR a useful resource for the plant research community.

ACKNOWLEDGEMENT

This work is supported by an Innovatie door Wetenschap en Technologie (IWT) grant for the Bioframe project and by an IUAP grant for the BioMaGNet project (ref. p6/25). We would also like to thank the anonymous reviewers for their constructive comments.

REFERENCES

Addo-Quaye C, Eshoo TW, Bartel DP, Axtell MJ (2008) Endogenous siRNA and miRNA targets identified by sequencing of the Arabidopsis degradome, *Curr Biol*, **18**, 758-762.

Allen E, Xie Z, Gustafson AM, Carrington JC (2005) microRNA-directed phasing during trans-acting siRNA biogenesis in plants, *Cell*, **121**, 207-221.

Alves, L., Jr., Niemeier, S., Hauenschild, A., Rehmsmeier, M. and Merkle, T. (2009) Comprehensive prediction of novel microRNA targets in Arabidopsis thaliana. *Nucleic Acids Res*, **37**, 4010-4021.

Brodersen P, Sakvarelidze-Achard L, Bruun-Rasmussen M, Dunoyer P, Yamamoto YY, et al. (2008) Widespread translational inhibition by plant miRNAs and siRNAs, *Science*, **320**, 1185-1190.

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.

Dugas DV, Bartel B (2008) Sucrose induction of Arabidopsis miR398 represses two Cu/Zn superoxide dismutases, *Plant Mol Biol*, **67**, 403-417.

Franco-Zorrilla, J.M., Valli, A., Todesco, M., Mateos, I., Puga, M.I., Rubio-Somoza, I., Leyva, A., Weigel, D., Garcia, J.A. and Paz-Ares, J. (2007) Target mimicry provides a new mechanism for regulation of microRNA activity, *Nat Genet*, **39**, 1033-1037.

German MA, Pillay M, Jeong DH, Hetawal A, Luo S, et al. (2008) Global identification of microRNA-target RNA pairs by parallel analysis of RNA ends, *Nat Biotechnol*, **26**, 941-946.

Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ (2008) miRBase: tools for microRNA genomics, *Nucleic Acids Res*, **36**, D154-158.

Hofacker I, Fontana W, Stadler P, Bonhoeffer S, Tacker M, et al. (1994) Fast Folding and Comparison of RNA Secondary Structures, *Monatsh Chem*, **125**, 167-188.

Kruger, J. and Rehmsmeier, M. (2006) RNAhybrid: microRNA target prediction easy, fast and flexible. *Nucleic Acids Res*, **34**, W451-454.

Moxon S, Schwach F, Dalmay T, Maclean D, Studholme DJ, et al. (2008) A toolkit for analysing large-scale plant small RNA datasets, *Bioinformatics*, **24**, 2252-2253.

Pearson W (2004) Finding protein and nucleotide similarities with FASTA, *Curr Protoc Bioinformatics*, **Chapter 3**, Unit3 9.

Rehmsmeier M, Steffen P, Hochsmann M, Giegerich R (2004) Fast and effective prediction of microRNA/target duplexes, *Rna*, **10**, 1507-1517.

Rhoades MW, Reinhart BJ, Lim LP, Burge CB, Bartel B, et al. (2002) Prediction of plant microRNA targets, *Cell*, **110**, 513-520.

Schwab R, Palatnik JF, Riester M, Schommer C, Schmid M, et al. (2005) Specific effects of microRNAs on the plant transcriptome, *Dev Cell*, **8**, 517-527.

Sunkar R, Zhu JK (2004) Novel and stress-regulated microRNAs and other small RNAs from Arabidopsis, *Plant Cell*, **16**, 2001-2019.

Voinnet, O. (2003) Origin, biogenesis, and activity of plant microRNAs, *Cell*, **136**, 669-687.

Zhang, Y. (2005) miRU: an automated plant miRNA target prediction server. *Nucleic Acids Res*, **33**, W701-704.