BIOINFORMATICS APPLICATIONS NOTE Vol. 22 no. 3 2006, pages 359–360
doi:10.1093/bioinformatics/bti802

Sequence analysis

Identification of plant microRNA homologs

Tobias Dezulian[1,*], Michael Remmert[1], Javier F. Palatnik[2], Detlef Weigel[2] and Daniel H. Huson[1]

[1]Center for Bioinformatics Tübingen, Tübingen University, Germany and [2]Max Planck Institute for Developmental Biology, Tübingen, Germany

Received on August 12, 2005; revised on November 10, 2005; accepted on November 24, 2005
Advance Access publication November 29, 2005
Associate Editor: Charlie Hodgman

ABSTRACT

Summary: MicroRNAs (miRNAs) are a recently discovered class of non-coding RNAs that regulate gene and protein expression in plants and animals. MiRNAs have so far been identified mostly by specific cloning of small RNA molecules, complemented by computational methods. We present a computational identification approach that is able to identify candidate miRNA homologs in any set of sequences, given a query miRNA. The approach is based on a sequence similarity search step followed by a set of structural filters.

Availability: microHARVESTER is offered as a web-service and additionally as source code upon request at http://www-ab.informatik.uni-tuebingen.de/software/microHARVESTER

Contact: dezulian@informatik.uni-tuebingen.de

MicroRNAs (miRNAs) are small RNAs 20–24 nt in length. They perform important regulatory roles in both plants and animals. The miRNA biogenesis and effector pathways share components with those for another class of small RNAs, short interfering RNAs (siRNAs), and both are currently under intense scrutiny (Susi *et al*., 2004). Biogenesis of miRNAs starts with the synthesis of a large primary transcript (Bartel, 2004; He and Hannon, 2004), which contains a double-stranded miRNA precursor that adopts a fold-back structure by complementary base pairing. In plants, the miRNA precursor is degraded in the nucleus by the RNAse III enzyme DICER-LIKE1, which releases a short RNA duplex. This duplex is formed by the miRNA along with the complementary fragment, called miRNA*, from the other arm of the precursor. The miRNA and the miRNA* are offset by 2 nt owing to the staggered cuts of DICER-LIKE1. Finally, mature miRNAs are selected from the RNA duplex and incorporated into RNA induced silencing complexes (RISCs), to which they provide sequence specificity.

MiRNAs recognize completely or partially complementary sequences in target mRNAs and guide them to cleavage or translational arrest. Animal miRNAs typically recognize several target sequences located in the 3′-UTR and inhibit their translation, whereas plant miRNAs usually recognize one motif in the coding region of their targets and affect their stability. It is thought that the better complementarity between plant miRNAs and their targets favors the latter mechanism. In plants, miRNAs regulate diverse genes and pathways, such as development, hormone signaling, stress response and trans-acting siRNAs (Allen *et al*., 2005).

MiRNA genes in plants are grouped into families that yield (almost) identical miRNAs. Currently, 43 families containing 513 miRNA genes across 7 plant species are listed in release 7.0 of the MicroRNA registry (Griffiths-Jones, 2004, http://microrna.sanger.ac.uk) responsible for name assignment of published miRNAs.

Here, we present an approach and implementation ('microHARVESTER') that can identify candidate miRNA homologs based on a query miRNA, with excellent sensitivity and specificity. The microHARVESTER takes advantage of the conservation pattern typical for miRNA genes: the (mature) miRNA is most conserved since its sequence is crucial for target-interaction; the miRNA* is less conserved but restricted by the need to extensively base-pair with the miRNA; the rest of the miRNA gene can be less conserved. Our approach uses a BLAST sequence similarity search to first generate a set of candidates which is then rigorously refined by a series of filters—exploiting structural features specific to plant miRNAs to achieve specificity. The output of the tool consists of a PDF overview document that is generated for each miRNA query. It presents candidate miRNA homologs along with figures of their predicted structure and a color-coded alignment.

Given a known miRNA (miRNA precursor sequence plus mature miRNA sequence) as input for our search we use the precursor as a query for a sequence similarity search against a set of sequences (e.g. a set of EST sequences or read from a new plant genome) to generate a set of candidate homologs. Since the (mature) miRNA sequence is very much conserved across large evolutionary distances (Axtell and Bartel, 2005), using BLAST (Altschul *et al*., 1997) with the very large *E*-value cutoff of 10 and minimal word size of 7, one can generate a hit for almost all miRNA homologs at the price of many false positives. In the first filter step, we discard those sequences of the candidate set whose aligned segments do not span most of the mature segment of the query. In a second filter step, we apply a modified Smith–Waterman pairwise alignment algorithm (Smith and Waterman, 1981) to precisely determine the mature sequence in the candidate precursor from the optimal alignment of the query mature sequence against the corresponding segment of the BLAST hit. We discard a candidate if the length of the mature sequences differs by >2 nt. In a third filter step, we predict the minimal free energy structure of the candidate sequence using RNAfold (Hofacker, 1994) and determine its putative miRNA* sequence. We discard a candidate if more than six nucleotides of its miRNA* are not predicted to form bonds with its mature miRNA (keeping in mind the 2 nt offset between miRNA

*To whom correspondence should be addressed.

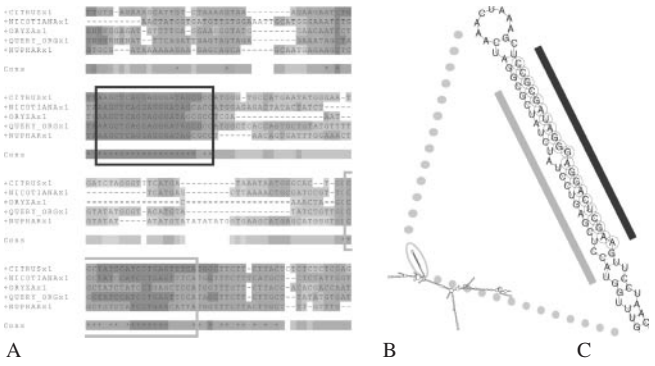

Fig. 1. (**A**) The multiple sequence alignment shows the reliability of each alignment position. Darker colors indicate better alignment scores. Dark and light frames mark the positions coding for the miRNA and miRNA*, respectively. (**B**) The minimal free energy structure for an EST harboring a miRNA homolog candidate is depicted; in the enlarged section (**C**), miRNA and miRNA* are marked on the right and left hand side, respectively.

and miRNA*). From a selection of all candidates that pass each filter we construct a multiple sequence alignment, using T-Coffee (Notredame *et al*., 2000), of a region that includes the miRNA, the miRNA* and the 'loop' sequence in between the miRNA and the miRNA*. The reliability of each position of this multiple alignment is visualized using a color scheme. An overview PDF document is generated, which contains this multiple sequence alignment. In addition, it provides for each putative miRNA homolog: a figure of its minimal free energy structure with the miRNA and the miRNA* highlighted in dark and light shades, respectively, along with its database accession (Fig. 1).

In order to assess sensitivity and specificity of this approach, we applied the microHARVESTER to the fully sequenced dicot *Arabidopsis thaliana* (Ath) genome using a set of query sequences from the monocot *Zea mays* (Zma). For each of the currently available (MicroRNA registry release 7.0) 18 miRNA families shared by Ath and Zma we selected one Zma miRNA gene at random. Using this query set, the microHARVESTER identified 67 of the 75 Ath miRNA genes of these families—at least one in each family—at the price of five false positives.

MicroHARVESTER is available as a web-service at www-ab.informatik.uni-tuebingen.de/software/microHARVESTER. Up to five miRNA queries may be submitted upon which a job id and URL will be issued and the resulting PDFs will be downloadable after job completion. Source code for the microHARVESTER is also available from the authors upon request. In order to run this standalone version on a standard linux operating system, additionally the following free software is needed: Java 1.5, NCBI BLAST, RNAfold, T-Coffee plus a standard LaTeX installation. Results can optionally be stored in a mySQL database. Note that when constructing the BLAST database, large input sequences are split into overlapping fragments for better retrieval efficiency.

MicroHARVESTER is able to identify plant miRNA homologs with good sensitivity and specificity in any set of sequences, for a given query miRNA. Using an EST database as the sequence pool offers the additional assurance that the predicted miRNA homologs are actually expressed (Zhang *et al*., 2005). Nevertheless, this approach has also proven useful on databases of genomic DNA.

Successful approaches for plant miRNA homolog identification have previously been described (Maher *et al*., 2004; Adai *et al*., 2005). However, microHARVESTER is the first such tool that is available through a web interface. It complements a very recently published animal miRNA homolog identification approach (Wang *et al*., 2005). In addition to the original purpose of miRNA homolog identification, microHARVESTER can be effectively used to screen candidate miRNA sets derived from comparative approaches to identify representatives of new miRNA families. In this setting, each candidate miRNA is used as the query and the number and divergence pattern of resulting putative homologs as well as their structure provides clues to the miRNA-likeness of the query.

Conflict of Interest: none declared.

REFERENCES

Adai,A., Johnson,C., Mlotshwa,S., Archer-Evans,S., Manocha,V., Vance,V. and Sundaresan,V. (2005) Computational prediction of miRNAs in *Arabidopsis thaliana*. *Genome Res*., **15**, 78–91.

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.

Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*., **25**, 3389–3402.

Axtell,M.J. and Bartel,D.P. (2005) Antiquity of MicroRNAs and their targets in land plants. *Plant Cell*, **17**, 1658–1673.

Bartel,D.P. (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, **116**, 281–297.

Griffiths-Jones,S. (2004) The microRNA Registry. *Nucleic Acids Res*., **32**, D109–D111.

He,L. and Hannon,G.J. (2004) MicroRNAs: small RNAs with a big role in gene regulation. *Nat. Rev. Genet*., **5**, 522–531.

Hofacker,I.L., Fontana,W., Stadler,P.F., Bonhoeffer,L.S., Tacker,M. and Schuster,P. (1994) Fast folding and comparison of RNA secondary structures. *Monatsh. Chem*., **125**, 167–188.

Maher,C., Timmermans,M., Stein,L. and Ware,D. (2004) Identifying MicroRNAs in Plant Genomes. *Proceedings of the 2004 IEEE Computational Systems Bioinformatics Conference (CSB 2004)* Stanford, CA, pp. 718–723.

Notredame,C., Higgins,D.G. and Heringa,J. (2000) T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol*., **302**, 205–217.

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

Susi,P., Hohkuri,M., Wahlroos,T. and Kilby,N.J. (2004) Characteristics of RNA silencing in plants: similarities and differences across kingdoms. *Plant Mol. Biol*., **54**, 157–174.

Wang,X., Zhang,J., Li,F., Gu,J., He,T., Zhang,X. and Li,Y. (2005) MicroRNA identification based on sequence and structure alignment. *Bioinformatics*, **21**, 3610–3614.

Zhang,B.H., Pan,X.P., Wang,Q.L., Cobb,G.P. and Anderson,T.A. (2005) Identification and characterization of new plant microRNAs using EST analysis. *Cell Res*., **15**, 336–360.