Computational identification of *Arabidopsis thaliana* microRNAs

Eric Bonnet, Jan Wuyts, Pierre Rouzé and Yves Van de Peer.

Department of Plant Systems Biology, Flanders Interuniversity Institute for Biotechnology (VIB), Ghent University, Technologie Park 927, B-9052 Gent, Belgium.

Contact: yves.vandepeer@psb.ugent.be

Introduction

MicroRNAs (miRNA) are an extensive class of tiny RNA molecules that are thought to regulate the expression of target genes via complementary base-pair interactions. Although the first miRNAs were discovered in the worm *Caenorhabditis elegans*, more than 200 miRNAs were recently found in diverse Eukaryotic organisms, most of the time by direct cloning methods. This approach is obviously biased in favor of the most abundant miRNAs. Some author already developed computational approaches to identify new miRNA genes in animals, using methods based on comparative genomics and also on characteristic patterns of the secondary structure of the miRNA precursor sequences. In this study, we present a genome-wide computational approach, called mirfinder, to detect new miRNA genes in the *Arabidopsis* genome, taking into account specific characteristics of the plant miRNAs.

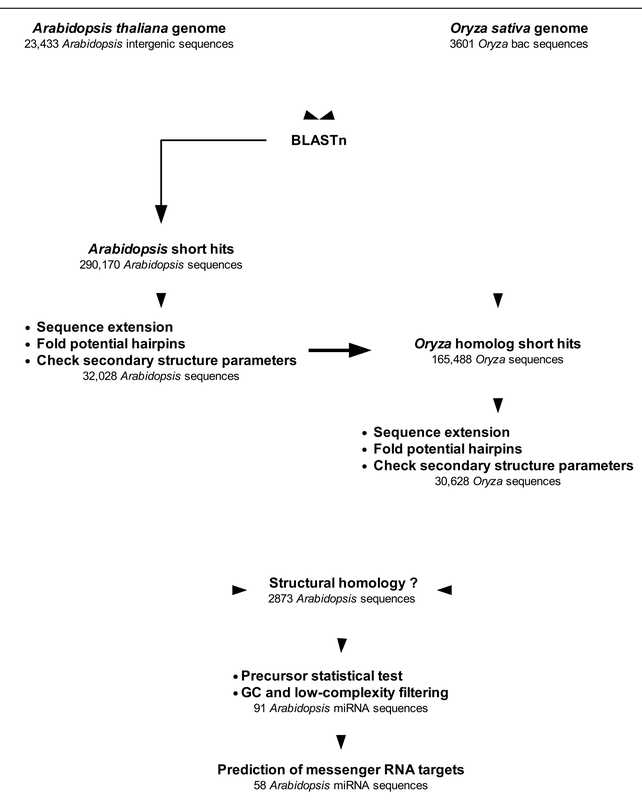
The *mirfinder* computational pipeline

The pipeline is based on **3 major rules** derived from a reference set constituted of 43 miRNAs previously described in the literature. All of them were characterized trhough experimental methods.

- The miRNA sequence is **conserved** between *Arabidopsis* and *Oryza* while the **rest of the precursor sequence** has diverged.
- Even though the precursor sequence has diverged, the ability of the precursor sequence to **form a similar stem-loop secondary structure** in both *Arabidopsis* and *Oryza* is conserved.
- For two miRNA orthologs, the miRNA sequence is always located on **the same arm** of the stem-loop secondary structures.

The pipeline is thus ordered as follows:

1. First look at short BLAST hits between Arabidopsis intergenic sequences and Oryza.
2. Extend those hits and look for potential stem-loop structures.
3. Check the folded structure of precursor candidates for compliance with a set of miRNA parameters.
4. Check if the sequences retained have a structural homolog in Oryza.
5. If the miRNA sequence is on the same arm (either 5' or 3') of the precursor both in Arabidopsis and Oryza, the sequence is considered as a valid candidate.
6. In order to avoid false-positives, candidates are filtered on GC content and entropy value for the miRNA sequence and also with a randomization statistical test on the precursor sequence.

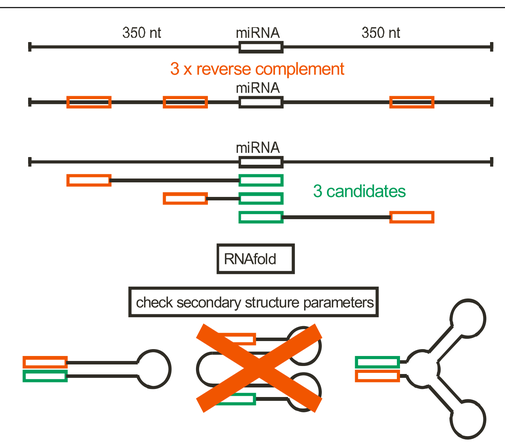


Selection of precursor candidates

All BLAST hits are extended upstream and downstream to a **window of 700 nt**. We used the **matcher** program from the EMBOSS package to scan for all close matches of the reverse complement of the initial BLAST hit (miRNA candidate) within the extended sequence.

This gives rise to a number of **potential fold-back structures** candidates that are subsequently folded by **RNAfold** of the Vienna package.

From these structures we extract a number of **parameters** that are compared to the equivalent parameters extracted from a "known good" reference set (see below).



Validation of precursor candidates

In order to select valid miRNA secondary structures, apart from the three general rules described above, **six characteristic features** of the secondary structure of miRNA were also derived from the reference set, of which one qualitative and five quantitative parameters. These parameters are (see also figure on the right):

(a) the miRNA should be part of a continuous helix

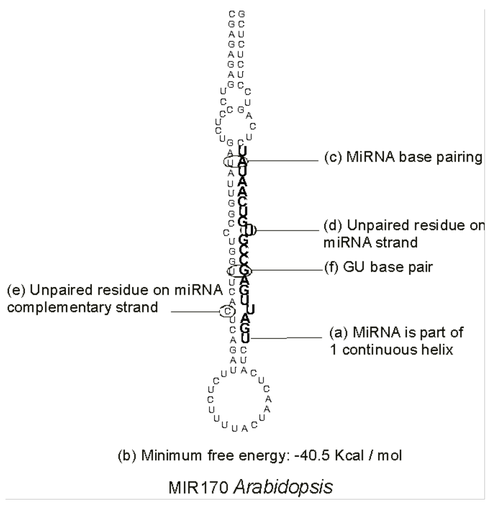
(b) the minimum free energy value should be less than –30Kcal/mol

(c) the minimum number of paired residues in the miRNA should be 15

(d) the maximum number of unpaired residues in the miRNA should be 5

(e) the maximum number of unpaired residues in the miRNA should be 5

(f) the maximum number of G.U pairs in the miRNA should be 5.



Results

After filtering of different RNA primary and secondary structure parameters we have identified **91 candidates** miRNA genes in *Arabidopsis* (see example in the figure on the right).

For example, the figure on the right shows 3 miRNAs that target the 3'UTR of gene At2g33770 (Ubiquitin-conjugating enzyme family) at 5 different places. For these 3 *Arabidopsis* miRNAs we found 4 homologs in *Oryza*.

The sensitivity of our approach is illustrated by the fact that we find back **6 of the 8 miRNA** genes which are known to be conserved between *Arabidopsis* and *Oryza*.

Plant miRNAs have the property of binding to their messenger RNA targets with a near-perfect complementarity. Using a program defined on this typical feature, we could predict a target in the *Arabidopsis* genome for **58 miRNAs**.

Half of the predicted targets belongs to the **transcription factor family**, which was already demonstrated in some studies.

Interestingly, the other targets are members of other cellular function like the **transport inhibitor response 1 (TIR1)**, which is involved in cellular communication/signal transduction.