

Identifying MicroRNAs in Plant Genomes

Christopher Maher Marja Timmermans Lincoln Stein Doreen Ware
Cold Spring Harbor Laboratory
maher@cshl.edu timmersma@cshl.edu lstein@cshl.edu ware@cshl.edu

Abstract

The ability to control gene expression during development in plants could be used for improving crop yields, resistance to disease, and environmental adaptability. It has been suggested that microRNAs, or miRNAs, control developmental processes such as meristem cell identity, organ polarity, and developmental timing by interfering with the expression of mRNAs. Our preliminary analysis focuses on the miR166 family since it has been shown to mediate repression of rolled-leaf1 (rdl1) in maize.

Based on maize sequences derived from degenerate primers, we computationally identified miR166b, miR166c, miR166d, and four more closely related putative maize precursors. Patscan, a pattern-matching program that allows RNA basepairing and mismatches, was used to identify functional elements in the putative hairpins. Each hairpin was further supported by their stable secondary structures determined with Mfold. Using this pattern matching approach we expanded the analysis for 19 Arabidopsis miRNA families in rice and maize.

1. Introduction

RNA-based gene regulation in plants is an evolutionarily conserved system involving two kinds of small RNAs [1]. Micro-RNAs (miRNAs) are 20 to 22-mers that usually basepair imperfectly to the coding region of a gene. Incompletely-characterized processes then induce cleavage or inhibit translation [2]. A second class of small RNAs are the small interfering RNAs (siRNAs), which are produced by cleavage of double stranded RNA (dsRNA) and can be broken down into 2 classes based on their length (20 to 22-mers and 24 to 26-mers). These RNAs, which match their mRNA targets exactly, induce cleavage-mediated degradation of their targets [3-5].

MiRNAs are the mature product of larger stem-loop precursor sequences transcribed from non-protein-coding genes that are processed by a ribonuclease III-like nuclease known as DICER-LIKE (dcl1) [2,6,7].

Due to the lack of compiled miRNA libraries in the cereal genomes, comparative analysis will play an important role in detection of maize miRNAs. The conservation of hairpin structures among multiple genomes suggests functional importance. Our focus is to take advantage of sequence conservation to develop a method for detecting miRNAs in cereal genomes.

In order to develop a high throughput method for detecting miRNAs, we needed to conduct a preliminary analysis on a known set. MiR166 has been proposed to mediate repression of *rolled-leaf1 (rdl1)* in maize and therefore we focused our preliminary analysis on the miR166 family in *Arabidopsis thaliana*, *Zea mays* (maize), and *Oryza Sativa* (rice) [8]. The protocol developed for identifying miR166 precursors was applied to 18 additional *Arabidopsis* miRNA families.

Juarez et al. demonstrated that there were multiple regions of conservation in the flanking sequences of hairpins within the *Arabidopsis* miR166 family that may be regulatory elements. Therefore, any putative miR166 precursors we identified could be expanded and then compared for regions of conservation.

2. Methods

2.1. MiR166 precursor identification

The first goal was to identify as many miR166 precursors within the maize methylation filtered [9] and hi-Cot sequences [10] through comparative analysis and pattern matching. Our initial dataset consisted of 2 *Arabidopsis* precursors (At_miR166a, At_miR166b), 2 rice precursors (Os_miR166b, Os_miR166d), and 4 maize precursors (Zm_miR166a-d). The *Arabidopsis* and *Oryza* hairpins were retrieved from the MicroRNA Registry, which is a database of published miRNAs assigned unique names [11]. The initial maize precursor sequences were produced from degenerate primers [8]. The maize gene enriched genomic sequences consisted of the TIGR maize methylation filtered and hi-Cot cluster sequences (Version 4.0) [12].

All of the precursors were then aligned to the

Table 1. MicroRNAs with perfect matches in plant genomes

	ARABIDOPSIS		RICE		MAIZE			
	MiR166a	MiR166b	miR166b	miR166d	miR166a	miR166b	miR166c	MiR166d
Arabidopsis								
Stem	1	2	5	5	5	5	5	8
Hairpin Stems	2	2	0	0	0	0	0	0
Rice								
Stem	3	3	2	2	4	7	6	12
Hairpin Stems	0	0	4	2	2	2	0	1
Maize								
Stem	3	3	4	2	8	6	4	5
Hairpin Stems	0	0	4	4	1	2	2	1

maize cluster sequences with blastn to detect sequence conservation (E-value < 0.001 with the Dust filter on) [13]. The results were divided into two sets shown in Table 2. The first set, ‘Stem’, contained all hairpins with one region of conservation, whereas the second set, ‘Hairpin Stems’, consisted of hairpins that displayed two regions of conservation within 300 nucleotides of one another, representing a putative hairpin. We believed that the regions of conservation would correspond to the stems of the hairpin, which are expected to be more evolutionary conserved than the loop. Upon manual inspection we identified each region of conservation as a stem of a miR166 precursor, as designated in the microRNA Registry.

At most, the *Arabidopsis* hairpins had sequence conservation with one stem of a rice or maize hairpin corresponding to the stem believed to result in the mature miRNA product [1]. The rice and maize hairpins demonstrated sequence conservation between the stems containing the miRNA and sometimes conservation between the stems opposite the miRNA. Although only the microRNA was detected in comparative analysis using *Arabidopsis*, the stem opposite the microRNA was sometimes detected between rice and maize comparisons. This would be expected since rice and maize are more recently diverged from one another.

All potential hairpin sequences were evaluated for their ability to form energetically favorable structures based on thermodynamic principles. Mfold was used to predict the optimal secondary structures for all putative hairpins along with their respective free energy of folding, G [14,15].

To distinguish between a conserved region that corresponds to a stem of a putative hairpin or a target we used the protocol shown in Figure 1. If a potential stem does not overlap a coding region, it can be discarded as a potential target, but still may be a stem in a putative hairpin whose opposing stem had weak homology with a hairpin from another species. In these cases we used Patscan, a pattern matching

program, to identify any hairpins within the maize cluster sequence from which the putative stem originated that had an opposing stem which was the reverse complement of the microRNA with up to 3 mismatches and 7 insertions [16]. All potential hairpins were then tested for ability to form stable secondary structures with MFold.

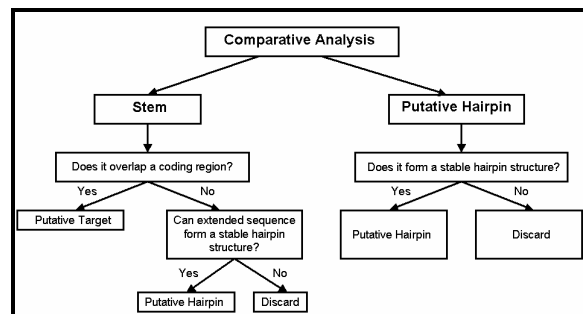


Figure 1. Flow chart of detecting putative hairpins

Table 2 shows all of the potential hairpins found throughout the maize cluster sequences. The term ‘novel hairpin’ indicates that it was unique from the four initial maize hairpins. The locations of three of the initial four maize precursors were identified, therefore allowing us to extract the flanking regions for further analysis. Four additional putative hairpins were detected, indicating that there are potentially at least 8 maize miR166 precursors.

Comparison of the secondary structures for miR166 precursors through *Arabidopsis*, rice, and maize shows that the mismatches and insertions between the stems of the hairpin are fairly consistent. Figure 2 shows the secondary structures from all three species in which the bulges are fairly similar, while there is a greater amount of variability within the loop sequence and length [17].

Table 2. Maize Precursors

Location	MiRNA	Start	End	Hairpin Length
AZM4_26017	Zm_miR166b	422	536	114
AZM4_51640	Novel	310	538	228
OGUER01TV	Novel	122	230	108
AZM4_70223	Zm_miR166c	3069	3182	113
AZM4_1320	Novel	1152	1260	108
AZM4_25993	Novel	309	520	211
AZM4_35871	Zm_miR166d	2529	2653	124

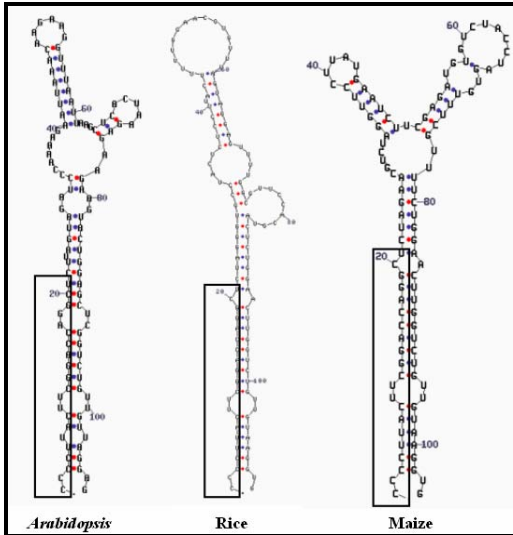


Figure 2. Precursor secondary structures determined by MFold

2.2 Hairpin extension analysis

It is believed that miRNAs are transcribed as long primary transcripts (pri-miRNAs) that are processed within the nucleus into stem-loop precursors (pre-miRNAs) followed by the cytoplasmic processing of pre-miRNAs into mature miRNAs [18]. Since the pri-miRNA extends beyond the hairpin, our goal was to identify sequence conservation between the flanking regions of the hairpins within the same family.

Previous studies have been limited in identifying pri-miRNA sequences because the maize genome is not fully sequenced. In order to maximize the length of the flanking regions in maize, we compared the whole cluster from which the precursor originated. The clusters from which the miR166 precursors were detected ranged in size from 1458 to 3435 nucleotides. Sequence alignment of the extended hairpin sequences was conducted within each species and between species using Pustell DNA matrix analysis within MacVector version 6.5.3 (hash value = 6; window size = 30).

The Pustell dot plots displayed conservation of

the stems between the hairpins both within and between species. For most of the comparisons within a species and between two species, minimal regions of conservation were detected outside of the hairpin. In the few occurrences where multiple regions of conservation were found outside of the hairpin, the two hairpins always belonged to the same species. This might indicate that the two precursors were recently duplicated from one another. Figure 3 is an example of two hairpins with many regions of sequence similarity flanking the precursor (enclosed within the square). Figure 4 shows only stem conservation between two maize hairpins (enclosed within the square), which is more typical within a species, and always seen between extended miR166 precursors of different species.

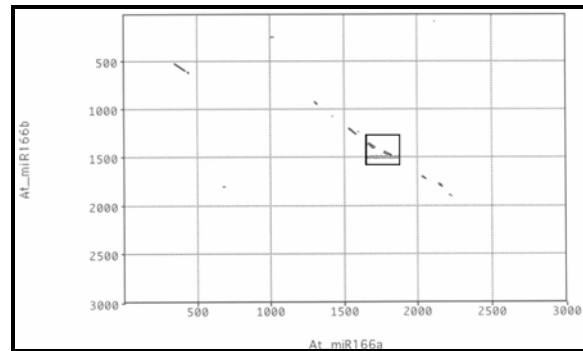


Figure 3. Pustell alignment of At_miR166a and At_miR166b

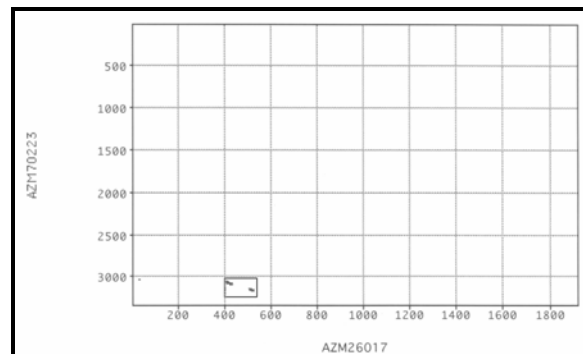


Figure 4. Pustell alignment of Zm_miR166b

and Zm_miR166d

Table 3. Frequency of perfect microRNA matches

Family	<i>Arabidopsis</i>	<i>Japonica</i>	<i>Indica</i>	<i>Zea Mays</i>
miR156	6	8	10	9
miR160	3	4	4	5
miR162	2	1	1	1
miR164	2	3	2	6
miR166	7	6	6	10
miR167	2	3	3	5
miR169	1	1	2	2
miR171	4	5	5	7
miR172	2	2	1	4

Table 4. Frequency of hairpins within plant genomes

Family	<i>Arabidopsis</i>	<i>Japonica</i>	<i>Indica</i>	<i>Zea Mays</i>	Mismatches	Insertions
miR156	6	9	11	10	1	2
miR160	3	4	4	6	3	1
miR162	2	1	1	1	3	0
miR164	2	3	3	4	5	0
miR166	8	7	7	11	6	0
miR167	2	3	3	5	2	2
miR169	1	1	1	2	3	0
miR171	0	0	0	1	2	0
miR172	3	2	1	4	4	0

The multiple regions of conservation between the two *Arabidopsis* sequences (At_miR166a and At_miR166b) were not detected when compared to other extended hairpin sequences. If these regions are regulatory elements, they are not conserved amongst all of the precursors within the family. Although the miRNAs are physically located on different chromosomes, and are therefore not tandem duplications, they reside near a large syntenic block as seen on the MIPS *Arabidopsis thaliana* database redundancy viewer (<http://mips.gsf.de/proj/thal/db/index.html>) (Figure 5). It is conceivable that these miRNAs could be recently duplicated genes and that is why we see the high level of conservation.

2.3 Analysis of 19 *Arabidopsis* miRNA families

The microRNA Registry lists 19 *Arabidopsis* miRNA families. Each of these families contains between one to eight precursors. Two of the families can produce slightly different miRNAs that differ by one nucleotide. Therefore our initial dataset consisted of 21 miRNAs.

By pattern matching the initial 21 miRNAs with Patscan, we identified all of the perfect matches

throughout multiple genomes [16]. In addition to the

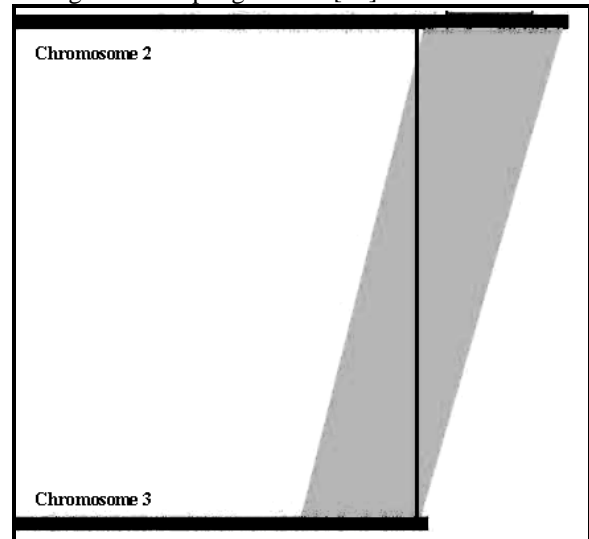


Figure 5. At_miR166b and At_miR166d reside near syntenic block between *Arabidopsis* chromosomes 2 and 3

maize cluster sequences used in the miR166 analysis we searched the TAIR *Arabidopsis thaliana* assembly (version 4.0), the TIGR *Oryza Sativa japonica* assembly (version 1.0), and the *Oryza Sativa indica*

shotgun sequences. Each perfect hit represents either a target, or a stem of a hairpin. Table 3 shows the 9, of the 19 *Arabidopsis* miRNA families, that have perfect matches in multiple species.

As shown with the miR166 family, there are common trends amongst the quantity and location of mismatches and insertions between the stems of a hairpin throughout the precursors of a miRNA family. Therefore we attempted to identify precursors that matched the hairpin pattern of that family. Table 4 shows the number of hairpins found within each family, and the maximum number of mismatches and insertions that we allowed. The mismatch and insertion thresholds were derived from the *Arabidopsis* family members under the assumption that the number of mismatches and insertions between the stems are fairly consistent between precursors of the same family as seen in miR166.

It has been observed that the *Arabidopsis* miR166 precursors not only have sequence conservation amongst the miRNA within each family member, but that the miRNA is physically located on the same stem within each precursor. This would indicate that they were derived from a common ancestor [5]. Although Patscan will search for a pattern on both strands, two sequence patterns needed to be designed to detect hairpins in which the microRNA is on the 5' stem or the 3' stem. Therefore we designed a sequence pattern containing the reverse complement of the desired hairpin. The complement of any resulting matches in the target sequence was the desired hairpin.

There are slightly more hairpins detected within both rice species than there were in *Arabidopsis*. The maize hairpins detected did not always outnumber the other species, despite having the largest genome, but this could be in part due to the incomplete nature of the maize genome.

Although we used a rule file to detect the miRNA in either the 5' or 3' stem of the hairpin, we expected only one of those two rule files would detect hairpins. In the instances in which a hairpin was detected with a miRNA in the wrong stem (miR156 and miR166), it has been due to two hairpins with the correct orientation being within 300 base pairs of one another. Therefore Patscan detected the reverse complement of the miRNA from one hairpin and the reverse complement of the opposite strand from the second hairpin forming a large hairpin on the opposite strand. These extra hairpins not only had very long loop regions, uncommon within the family, but also produced very different secondary structures than the remaining members of the family.

Within the miR156 family, the two hairpins that were physically located near one another, therefore allowing the detection of a false hairpin, were found in both rice species and maize but not in *Arabidopsis*.

Therefore this tandem duplication may have occurred since the monocot-dicot divergence.

In comparison to other Clustalw alignments of loop regions between precursors of the miR166 family, the loop regions of these two very closely related hairpins are fairly well conserved in both sequence and length throughout both rice species and maize [18]. Figure 6 shows the Clustalw alignments of the loop region for each of the precursors in this recently duplicated pair.

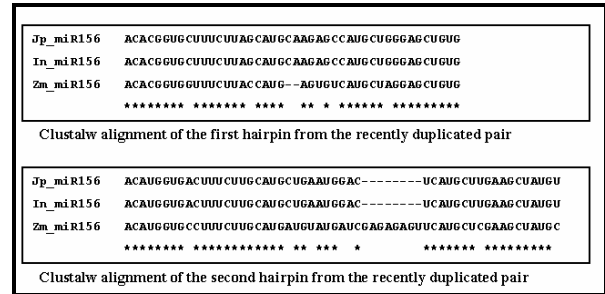


Figure 6. Clustalw alignment of the loop region

3. Discussion

Using a pattern matching approach, we have developed a method for detecting miRNA precursors within cereal genomes. All of our analysis to date used the *Arabidopsis* miRNA as a starting point, relying on the evolutionary constraints placed on the miRNA to detect precursors within other species.

Detection of miRNAs using blastn demonstrated that more distantly related plant species could still display conservation in the stem resulting in the miRNA while more closely related species can show some conservation in both stems of the hairpin. Although sequence alignment tools can pick up miRNA conservation between monocots and dicots, pattern matching has provided a more accurate method for precursor detection.

Due to the miR166 hairpin similarity between species, we can easily modify the mismatch and insertion criteria for each additional family. For every perfect miRNA hit found for each family throughout each genome, a corresponding hairpin could be detected (Table 3 and Table 4) with family specific pattern matching criteria. If this analysis was expanded for detecting novel precursor families, the mismatch and insertion criteria would be looser, to avoid being biased towards one family, but can remain constant throughout the species with the expectation of detecting most of the precursors.

It can be seen in *Arabidopsis* that precursors within the same family may be recently duplicated

from one another. By understanding how these miRNA families have evolved, we hope to gain some insight into their functionality.

Given the imperfect base pairing between the mRNA and miRNA, we plan on identifying potential targets for each precursor using a pattern matching approach. Computationally predicted targets will then be experimentally verified.

4. References

[1] Reinhart, B.J., Weinstein, E.G., Rhoades, M.W., Bartel, B. and Bartel, D.P. "MicroRNAs in plants." *Genes Dev.* 2002 16 pp. 1616–1626.

[2] Bartel, B.P., "MicroRNAs: genomics, biogenesis, mechanism, and function." *Cell.* 2004 Jan 23;116(2):281-97.

[3] Tijsterman M, Ketting RF, Plasterk RH. "The genetics of RNA silencing." *Annu Rev Genet.* 2002;36:489-519. Epub Jun 11.

[4] Tang G, Reinhart BJ, Bartel DP, Zamore PD. "A biochemical framework for RNA silencing in plants." *Genes Dev.* 2003 Jan 1;17(1):49-63

[5] C. Llave, K. Kasschau, M. Rector and J. Carrington. "Endogenous and silencing-associated small RNAs in plants." *Plant Cell* 2002 14 pp. 1605–1619.

[6] Bartel, B. and Bartel, D.P.. "MicroRNAs: At the root of plant development?." *Plant Physiol.* 2003 132, pp. 709–717.

[7] Hannon, G.J., "RNA interference." *Nature.* 2002 Jul 11;418(6894):244-51.

[8] Juarez, M. et al., "microRNA-mediated repression of *rolled leaf1* specifies maize leaf polarity." *Nature.* 2004 March 4;428:84-8.

[9] Rabinowicz PD, Schutz K, Dedhia N, Yordan C, Parnell LD, Stein L, McCombie WR, Martienssen RA. "Differential methylation of genes and retrotransposons facilitates shotgun sequencing of the maize genome." *Nat*

Genet. 1999 Nov;23(3):305-8

[10] Yuan Y, SanMiguel PJ, Bennetzen JL. "High-Cot sequence analysis of the maize genome." *Plant J.* 2003 Apr;34(2):249-55.

[11] Whitelaw CA. et al. "Enrichment of gene-coding sequences in maize by genome filtration." *Science.* 2003 Dec 19;302(5653):2118-20.

[12] Griffiths-Jones S. "The microRNA Registry." *Nucleic Acids Res.* 2004 Jan 1;32.

[13] Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. "Basic local alignment search tool." *J. Mol. Biol.* 1990 215:403-410.

[14] Turner D.H. et al.. Expanded Sequence Dependence of Thermodynamic Parameters Provides Robust Prediction of RNA Secondary Structure. *J. Mol. Biol.* 1999 288:910-940.

[15] Turner D.H. et al.. "Algorithms and Thermodynamics for RNA Secondary Structure Prediction: A Practical Guide. In RNA Biochemistry and Biotechnology, J. Barciszewski & B.F.C. Clark, eds., 1999 NATO ASI Series, Kluwer Academic Publishers.

[16] Dsouza M, Larsen N, Overbeek R., "Searching for patterns in genomic data", *Trends Genet.* 1997 Dec;13(12):497-8.

[17] Bartel, D.P. et al. MicroRNAs in plants. *Genes and Development.* 2002 Vol. 16(13):1616-26.

[18] Lee, Y. et al., "The nuclear RNase III Drosha initiates microRNA processing." *Nature.* 2003 Sep 25;425(6956):415-9.

[19] Higgins D., Thompson J., Gibson T. Thompson J.D., Higgins D.G., Gibson T.J. "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice." *Nucleic Acids Res.* 1994 22:4673-4680.