SAGE-Hindawi Access to Research
Journal of Nucleic Acids
Volume 2010, Article ID 495904, 10 pages
doi:10.4061/2010/495904

Research Article

NOVOMIR: De Novo Prediction of MicroRNA-Coding Regions in a Single Plant-Genome

Jan-Hendrik Teune and Gerhard Steger

Institut für Physikalische Biologie, Universitätsstr. 1, Heinrich-Heine-Universität Düsseldorf, 40225 Düsseldorf, Germany

Correspondence should be addressed to Gerhard Steger, steger@biophys.uni-duesseldorf.de

Received 25 March 2010; Revised 10 June 2010; Accepted 29 June 2010

Academic Editor: Ben Berkhout

MicroRNAs (miRNA) are small regulatory, noncoding RNA molecules that are transcribed as primary miRNAs (pri-miRNA) from eukaryotic genomes. At least in plants, their regulatory activity is mediated through base-pairing with protein-coding messenger RNAs (mRNA) followed by mRNA degradation or translation repression. We describe NOVOMIR, a program for the identification of miRNA genes in plant genomes. It uses a series of filter steps and a statistical model to discriminate a pre-miRNA from other RNAs and does rely neither on prior knowledge of a miRNA target nor on comparative genomics. The sensitivity and specificity of NOVOMIR for detection of premiRNAs from *Arabidopsis thaliana* is ~0.83 and ~0.99, respectively. Plant pre-miRNAs are more heterogeneous with respect to size and structure than animal pre-miRNAs. Despite these difficulties, NOVOMIR is well suited to perform searches for pre-miRNAs on a genomic scale. NOVOMIR is written in Perl and relies on two additional, free programs for prediction of RNA secondary structure (RNALFOLD, RNASHAPES).

1. Introduction

MicroRNAs (miRNAs) are genome-encoded single-stranded RNA molecules of ~22 nt in length, which play a significant role in regulation of gene expression in eukaryotes. Many details on biogenesis and interactions of miRNAs are known (see recent reviews, e.g., [1, 2]). Briefly, miRNAs can be encoded by miRNA genes, but also be generated from different RNA transcripts (e.g., from introns of protein-coding genes). Plant and animal miRNAs differ to some extent with respect to biogenesis and structural characteristics but also in their mode of action. In plants, most if not all miRNAs are transcribed from genes by RNA-dependent RNA polymerase II (polII) into primary transcripts called pri-miRNA; these transcripts fold into (possibly imperfect) stem-loop structures. From the pri-miRNA Dicer-like (DCL) enzymes process the stem-loop structure (pre-miRNA), which is usually longer (~130 nt; see below) than nonplant pre-miRNA (~86 nt), and finally a miRNA/miRNA* duplex. In the cytoplasm, the miRNA is incorporated into the RNA-induced silencing complex (RISC), and base-pairing of the miRNA with complementary messenger RNA (mRNA) regions leads to mRNA degradation or to inhibition of mRNA translation. Most plant miRNAs base-pair with their respective target mRNAs in the coding region with perfect or near-perfect complementarity leading to cleavage (and degradation) of the mRNAs; animal miRNAs usually base-pair with 3′ untranslated regions through imperfect complementarity leading to translation repression.

Finding of miRNA genes either needs costly experimental approaches—for example, genetics, which led to the detection of the first animal miRNAs [3, 4], cloning and sequencing of cDNA, or deep sequencing—or computational prediction methods, which facilitate subsequent experimental verification or falsification. The different properties of miRNAs in plants and animals gave rise to different computational approaches (for reviews see [5–7]). Most of these tools, however, rely on the following features: the miRNA resides in a stem-loop structure, which possess a high thermodynamic stability and does not contain large internal loops or asymmetric bulges at least in the region of the mature miRNA [8]. In addition, many tools take into account a phylogenetic conservation of the pre-miRNA structure and miRNA sequence, which limits the chance to

detect non-conserved, evolutionary new miRNA genes. For example, Dezulian et al. [9] identify plant miRNA homologs in a set of sequences, given a query miRNA, by a sequence similarity search step and a set of structural filters; Pfeffer et al. [10] identify DNA-viral pre-miRNAs, which show neither detectable conservation to other viral pre-miRNAs nor to host pre-miRNAs, by a search for stable stem-loops and scoring of these according to free energy of folding, base composition, and number of base pairs; Wang et al. [11] as well as Jones-Rhoades and Bartel [12] search for putative miRNA/miRNA* complexes in the intergenic regions of *Arabidopsis thaliana* and filter these according to GC content, mismatches in the stem, conservation in the rice genome, and the characteristic stem-loop structure.

To our knowledge, the only tools for *de novo* prediction of pre-miRNAs in plants are HHMMiR [13] and triplet-SVM [14]. HHMMiR calculates first the mfe structure of sequence regions (using RNAfold in a scanning window approach with window length of less than 500 nt), extracts stem-loops that possess at least 10 base pairs, a minimum length of 50 nt, a loop of less than 20 nt and no multiloop(s), and finally classifies *via* a hierarchical hidden Markov model (HHMM). The sensitivity of HHMMiR is published to be 0.865 for *Oryza sativa* (96 sequences taken from miRBase 5) and 0.973 for *A. thaliana* (75 sequences). triplet-SVM calculates by RNAfold the mfe structure of sequences, rejects those with junction(s), too few base pairs, and a high free energy (i.e., low structural stability), parses the remaining structures in "triplets" (type of nucleotide plus paired or unpaired state of the nucleotide and its two neighbors), and finally classifies these features with a support vector machine (SVM). The sensitivity of triplet-SVM is published to be 0.948 for *Oryza sativa* and 0.92 for *A. thaliana* using the same sequences from miRBase 5 as in the test with HHMMiR.

In the following, we describe our tool, called novoMIR, to detect pre-miRNA and miRNA/miRNA* sequences in a plant genome. For this purpose novoMIR uses a series of filter steps, similar to those mentioned above, followed by a statistical model to discriminate a pre-miRNA from all other RNAs and by another statistical model to locate the miRNA/miRNA* complex in a putative pre-miRNA. Thresholds and statistical values are learned from sets of true positive sequences (plant pre-miRNAs taken from miRBase; [15]) and true non-miRNA sequences (tRNAs, 5 S rRNA, 5.8 S rRNA, mRNAs, etc.). For detection, novoMIR relies neither on comparative genomics nor on prior knowledge of a miRNA target; thus novoMIR allows for searches in single plant genomes as well as in viral or viroid genomes.

2. Methods

2.1. Features of Plant Pre-miRNA. Sequences of plant pre-miRNAs were obtained from different versions of miRBase [15, 16]: version 10.0 contains 1,247 sequences; the recent version 14 contains 2,030 sequences. The mean and median length of plant sequences are about (150 ± 73) nt and 130 nt, respectively (see Figure S1 in Supplementary Material available online at doi:10.4061/2010/495904); the shortest pre-miRNA is 54 nt in length (miRBase ID: gma-MIR2107) and the longest is 932 nt (cre-MIR916). The mean and median length of nonplant sequences are about (88 ± 14) nt and 86 nt, respectively; the shortest pre-miRNA is 44 nt in length (hsa-mir-1973) and the longest is 215 nt (dme-mir-997). That is, most plant pre-miRNAs are longer than animal pre-miRNAs and their size range is more diverse. The sequences of pre-miRNAs and mature miRNAs are slightly enriched in U [17] and U plus G, respectively (see Figure S2). The four nucleotides are not equally distributed at each position along the miRNA sequences (see Figure S3): for example, a U is the preferred 5′ nucleotide ($f_{1,\mathrm{U}} = 0.65$), a G on position 8 ($f_{8,\mathrm{G}} = 0.44$), and a C on position 19 ($f_{19,\mathrm{C}} = 0.52$). The minimum free energy $\Delta G^0_{37^\circ\mathrm{C}}$ of the secondary structures of pre-miRNAs, as calculated by RNAfold [18] using default parameters, is in a wide range due to the different lengths L and G+C contents f_{GC} of the sequences (see Figure S4); normalization of $\Delta G^0_{37^\circ\mathrm{C}}$ to length and f_{GC} [17] results in $\Delta G^0_{37^\circ\mathrm{C}}/L = (-0.45 \pm 0.12)$ kcal/mol/nt and $\Delta G^0_{37^\circ\mathrm{C}}/L/f_{\mathrm{GC}} = (-1.02 \pm 0.26)$ kcal/mol/nt; the latter value is significantly lower than that of other RNA according to Zhang et al. [17].

2.2. Training Data. We used the 184 pre-miRNAs and mature miRNAs of *A. thaliana* as listed in miRBase version 10 as the true-positive data set for establishing all thresholds and parameters of novoMIR. Sequences containing nucleotides other than A, C, G, U(T) were discarded. For evaluation of sensitivity we used in addition the plant pre-miRNAs and mature miRNAs from miRBase version 14 (190 from *A. thaliana* and 1,853 from other plants). The sensitivity of novoMIR was nearly identical for both data sets (and also with sequences from version 14 minus those from version 10; see supplemental Table S1); thus we refrained from training with different data sets.

2.3. Test Data. As the true-negative data set, we assembled RNA sets from the following sources:

(i) 710 mRNA sequences randomly selected from *A. thaliana*

(ii) 631 tRNA sequences from *A. thaliana*

(iii) 63 5.8 S rRNA sequences from Rfam version 7.0 [19]

(iv) 602 5 S rRNA sequences from Rfam version 7.0

(v) one randomly selected RNA sequence from each of the 455 noncoding RNA families from Rfam version 7.0 (except miRNA families);

(vi) 2,760 shuffled pre-miRNA sequences (each of the 184 *A. thaliana* sequences from miRBase 10 was shuffled 5 times using shuffle [20] preserving (a) the mononucleotide content, (b) mono- and dinucleotide content, and (c) mononucleotide content in a window of 20 nt, resp.)

(vii) repetitive genomic elements from *A. thaliana* from the RepeatMasker library [21] (in total 134,000 nt)

(viii) 8,000 pseudohairpin sequences from *Homo sapiens* [22]

(ix) 10,000 pseudohairpin sequences from *A. thaliana*; these were selected using RNALfold from the TAIR cDNA library [23] to have a minimum stem-loop length of 50 nt in a base pair span of 400 nt

(x) $10 \times 5,000$ sequences of a length between 80 and 800 nts randomly selected from the five chromosomes of *A. thaliana*.

2.4. Availability and Requirements. novoMIR is written in Perl and was tested under Linux. It relies on RNAshapes [24, 25] and RNALfold [26] (which is part of the Vienna RNA package [18]) for secondary structure calculations. RNALfold finds subsequences of a long RNA sequence that fold into locally stable (i.e., thermodynamically favorable) RNA secondary structures; the computational effort is $\mathcal{O}(NL^2)$ with length N of the long RNA sequence and maximal base-pair separation L of the subsequences. For an RNA sequence, RNAshapes computes shapes, which are classes of similar secondary structures, and a representative structure ("shrep") of minimal free energy within each shape.

3. Algorithm

In the following, we describe the workflow of novoMIR (see supplemental Figure S5).

(1) A typical plant pre-miRNA consists of a relatively short sequence (with median length ~130 nt and mean length ~ 150 ± 73 nt) that is able to fold into a stable stem-loop structure. Thus, we search in the genomic sequence for subsequences with locally stable secondary structure(s) *via* RNALfold. In case the genomic sequence is longer than 1000 nt, we subdivide it into 1000 nt fragments overlapping by 400 nt. We choose a maximal base pair separation $L =$ 400 nt. This limit excludes only a few exceptionally long pre-miRNAs; that is, only 8 of 1356 plant pre-miRNA sequences in miRBase 10 and 14 of 2030 in miRBase 14, respectively, are dismissed due to this restriction for the sake of a fast first step. From the output of RNALfold, the five subsequences with best locally stable structures are treated further as individual sequences.

(2) The original sequence (with length ≤1000 nt) or a subsequence (with length ≤400 nt) selected by RNALfold is discarded if the sequence has a base composition not typical for pre-miRNAs; that is, the sequence is only retained if the fraction of each nucleotide is above 0.1. This filter rejects 9 and 21 plant pre-miRNA sequences from miRBase 10 and 14, respectively.

(3) RNAshapes is used to predict the thermodynamically optimal secondary structure (minimum free energy (mfe) structure with ΔG^0_{mfe}) and the optimal secondary structure of up to three shapes with energies less favorable than that of the mfe shape class by 0.1 kcal/mol. The shapes have to differ in their nesting pattern for all loop types but positions of unpaired regions are not of relevance (RNAshapes's option `-t 3`). In general, it is assumed that the mfe structure of pre-miRNAs is the conformation adequate for further processing by Dicer. In our case, however, we do not know the true 5′ and 3′ ends; thus, the unrelated termini of the respective sequence, which do not belong to the true pre-miRNA, might cause the pre-miRNA structure to be thermodynamically suboptimal. Moreover, the restriction by RNAshapes to the shrep prediction avoids prediction (and further processing) of the immense number of suboptimal structures.

(4) Any sequence that is not able to fold into a structure (as predicted in step (3)) with $\Delta G^0_{37^\circ C}/L/f_{\text{GC}} \leq -0.75$ kcal/mol/nt is rejected.

(5) Next, each retained secondary structure is reformatted from the bracket-dot notation used by RNAshapes into an alignment-like format [27] (for an example see Figure 1), which eases handling during the following steps: at each multiloop, the structure is divided into the respective stem-loop structures, which are separately processed further; 5′ and 3′ dangling ends are removed; a hairpin loop is removed; and asymmetric loops are made symmetric by introduction of gap symbols. Afterwards each (sub)structure consists of the following states: base pairs (match states M symbolized by $\begin{smallmatrix}+\\+\end{smallmatrix}$), loop "pairs" (mismatched states N, $\begin{smallmatrix}-\\-\end{smallmatrix}$), and insertion (I) and deletion (D) states ($\begin{smallmatrix}-\\|\end{smallmatrix}$ and $\begin{smallmatrix}|\\-\end{smallmatrix}$, resp.).

(6) A stem-loop shorter than 30 states in the alignment-like format is deleted. For efficiency of this filter, see Figure 2(a).

(7) Next, a window of length 25 states is moved (in steps of (1) state) along the structure in the alignment-like format, and the fraction of base-paired states is determined for each window. A stem-loop is deleted unless at least a mean fraction of 0.65 base-paired states is present in five different windows, which might overlap. For efficiency of this filter see Figure 2(b).

(8) A stem-loop is deleted if it does not contain a helix with at least 8 consecutive base pairs. For efficiency of this filter see Figure 2(c).

(9) A stem-loop is deleted if the ratio of its sequence length (as predicted by RNALfold) and the length of the stem-loop in the alignment-like format is above 6; that is, the structure contains too many junctions and/or large, unstructured hairpin loops. For efficiency of this filter see Figure 2(d).

(10) If a sequence (and structure) remains after the filter steps, novoMIR decides on its possibility to be a pre-miRNA using a paired Hidden-Markov model identical to that described by Nam et al. [27]. Briefly,

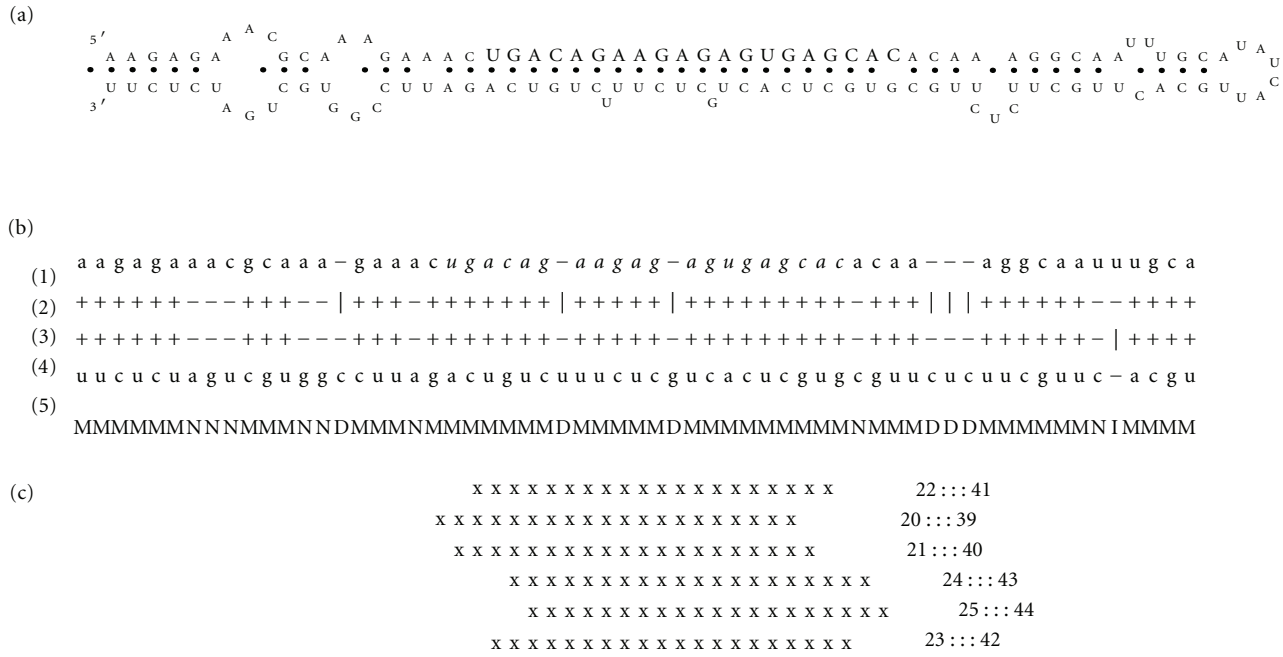


Figure 1: Example of a reformatted pre-miRNA structure and predicted localization of a miRNA/miRNA* complex in the pre-miRNA structure. (a) Secondary structure of the pre-miRNA of ath-mir156a in standard representation. The mature miRNA is shown in larger italic characters. (b) Section of the pre-miRNA in an alignment-like format. First and fourth line show the sequence except the sequence region of the hairpin loop; lines 2 and 3 describe the state of the opposing nucleotides: a base pair is marked by "+", nucleotides of an internal loop by "−" in both lines, and nucleotides that are part of an asymmetrical internal or a bulge loop by "|". In the fifth line a base pair is marked by "M", a non-pair by "N", a deletion in the top strand by "D", and an insertion in the top strand by "I". (c) Positions predicted as miRNA/miRNA* complexes are marked by "x" and their relative positions in the sequence.

the joint probability $P(x, \pi)$ of an observed sequence x and a state sequence π is

$$P(x,\pi) = T_{0\pi_1} \prod_{i=1}^{L} E_{\pi_i}(x_i) T_{\pi_i,\pi_{i+1}}, \tag{1}$$

with transition probabilities $T_{kl} = P(\pi_i = l \mid \pi_{i-1} = k)$ between the four states $k, l \in \{\mathrm{M}, \mathrm{N}, \mathrm{I}, \mathrm{D}\}$, emission probabilities $E_k(b) = P(x_i = b \mid \pi_i = k)$ of the different nucleotide and gap pairs b, window size $L = 21$, and the probability of starting in state k defined as $T_{0\pi_1}$. In contrast to Nam et al. [27], we use four hidden states (`is_miRNA`, `is_miRNA` → `is_not_miRNA`,`is_not_miRNA` → `is_miRNA`, `is_not_miRNA`; see Figure S6). For the decision that the sequence is a pre-miRNA or not, the values for the $j \in \{$`is_miRNA`, `is_not_miRNA`$\}$ states are normalized and summed up

$$P_j = \sum_{i=1}^{L} \frac{E_{\pi_{i,j}}\left(x_{i,j}\right) T_{\pi_{i,j},\pi_{i+1,j}}}{\sum_{j=1}^{4} E_{\pi_{i,j}}\left(x_{i,j}\right) T_{\pi_{i,j},\pi_{i+1,j}}}. \tag{2}$$

The squared ratio

$$R_1 = \left(\frac{P_{\texttt{is\_miRNA}}}{P_{\texttt{is\_not\_miRNA}}}\right)^2, \tag{3}$$

as well as the mean of the nine highest values of the difference

$$R_2 = \max \sum_{k=l}^{l+20} P_{\texttt{is\_miRNA}} - P_{\texttt{is\_not\_miRNA}}, \tag{4}$$

are compared to thresholds for the pre-miRNA decision.

(11) In case of a positive decision in the previous step, the values that lead to the six highest values of R_2 are predicted as positions of probable miRNA/miRNA* duplices (see Figure 1(c)).

4. Results and Discussion

Our program novoMIR uses a set of heuristic filters and a statistical model to discriminate a miRNA precursor from all other RNAs (see Figure S5). The data for this model are collected based on a set of true positive sequences (miRNA precursors from *A. thaliana* as in miRBase 10) and a set of true non-miRNA sequences (for details, see Section 2.3).

4.1. Test on miRBase *Version 14.* All thresholds for the filter steps and the probabilities for the Hidden-Markov model were selected on the basis of "receiver operating characteristic" (ROC) curves like those shown in Figure 2. For these, the set of true positive *A. thaliana* pre-miRNA

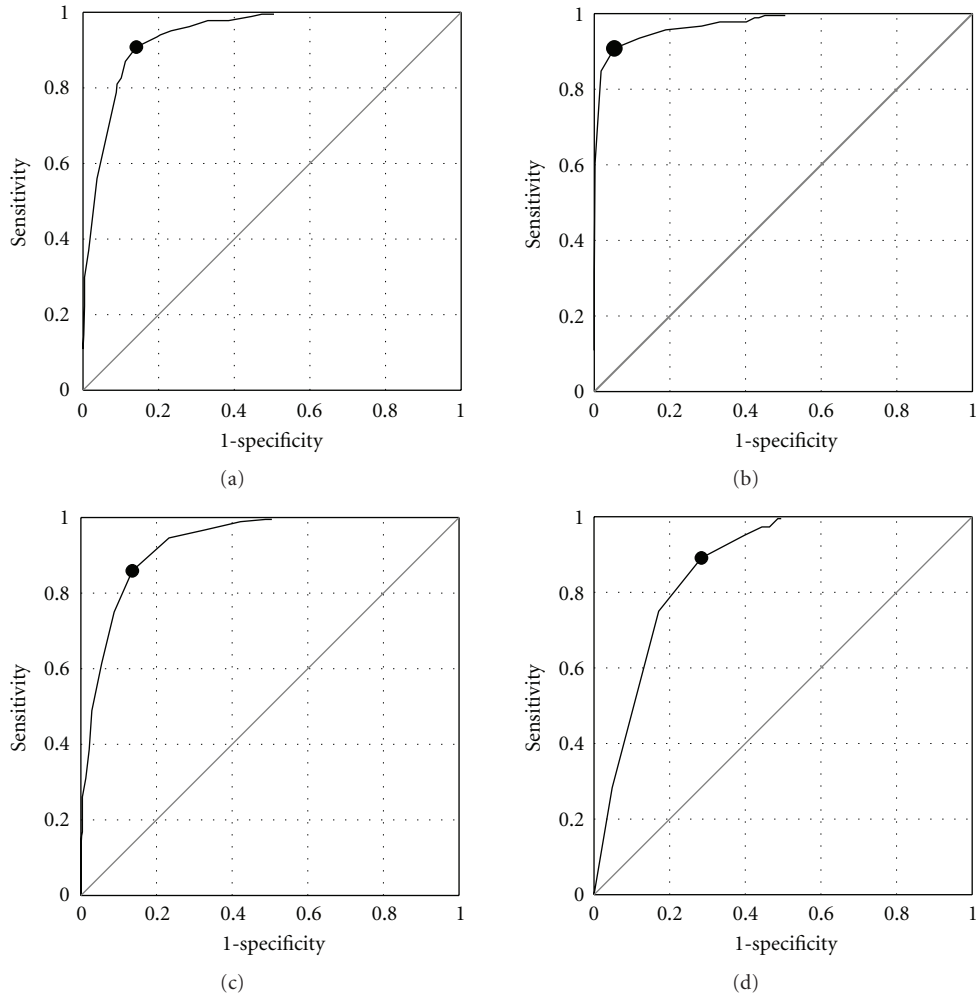


Figure 2: Efficiency of filtering steps and paired HMM. Receiver operating characteristic (ROC) curves for filters on (a) minimum length of stem-loop region, (b) fraction of base pairs in a sliding window of 25 states, (c) number of consecutive base pairs, and (d) ratio of sequence and stem-loop length. The area under the curve (AUC) is (a) 0.94, (b) 0.97, (c) 0.93, and (d) 0.87. The dots (at max(sensitivity + specificity − 1)) denote the value pair of sensitivity and false positive rate that optimally discriminates between miRNA and non-miRNA sequences. The data set consisted of all plant miRNA sequences from miRBase 10 and 455 non-miRNA sequences from Rfam 7.

sequences was taken from miRBase version 10. Sensitivity values for the enlarged set of pre-miRNAs from miRBase version 14 (190 *A. thaliana* and 1,840 sequences from other plants) are compared to those obtained from miRBase version 10 (184 *A. thaliana* and 1,063 sequences from other plants) in Table 1. The sensitivity values of novoMIR for *A. thaliana* pre-miRNA sequences of both miRBase versions are very close to each other (0.837 and 0.832, resp.). The values for all plant pre-miRNA sequences are slightly lower (0.791 and 0.792, resp.), but show no clear trend that sequences of miRBase 14 (not present in miRBase 10) are different from those of miRBase 10 or that sequences from a certain taxonomic group might be different from those of others (see supplemental Table S1).

The sensitivity of novoMIR in predicting the position of the miRNA/miRNA* complex is also high (0.73 for *A. thaliana* and 0.82 for all plants; see Table 1). For this, a position is counted as correctly predicted if it matches exactly the annotated mature miRNA or overlaps by five or fewer nucleotides.

4.2. Comparison with Other Tools. We tested HHMMiR [13] and triplet-SVM [14] for sensitivity with the sequences from miRBase 10 and 14 (see Table 1). Their sensitivity is at maximum 0.15 and 0.45, respectively. The filtering steps of both tools reject already many sequences (HHMMiR more than 80% and triplet-SVM more than 22%). For the sequences remaining after the filtering steps, the sensitivity of the HHMM and SVM is at maximum 0.79 and 0.60, respectively, which is also lower than that of novoMIR with a sensitivity of at least 0.80 (using all filter steps).

4.3. Tests on Specificity. We assembled different data sets to test the specificity of novoMIR. These data sets should not contain any true (pre-)miRNA. For example, we used well-annotated RNAs (mRNA, noncoding RNA) and sets of

Table 1: Sensitivity of novoMIR, triplet-SVM [14], and HHMMiR [13] in pre-miRNA prediction for different versions of miRBase. The row "14–10" shows values for sequences from miRBase 14 which are not present in miRBase 10.

miRBase version	# sequences		Sensitivity[1] novoMIR[2] pre-miRNA	novoMIR[2] miRNA/miRNA*	HHMMiR[3] pre-miRNA		triplet-SVM[3] pre-miRNA	
10	184	*A. th.*	0.84	0.73	0.15	0.75	0.45	0.60
14	190	*A. th.*	0.83	0.75	0.10	0.79	0.44	0.59
10	1247	plant	0.79	0.82	0.04	0.58	0.39	0.51
14	2030	plant	0.79	0.83	0.04	0.64	0.38	0.50
14–10	788	plant	0.80	—	0.04	0.73	0.38	0.48

[1] Sensitivity is calculated as $TP/(TP + FN)$.
[2] Note that novoMIR's thresholds and probabilities were learned only from *A. thaliana* sequences in miRBase version 10.
[3] The left column gives sensitivity for all sequences; the right column gives sensitivity for those sequences left after the preprocessing step(s) of HHMMiR and triplet-SVM, respectively.

"pseudohairpins" from *H. sapiens* and *A. thaliana*. Similarly, the chance is negligible that the data set of $10 \times 5,000$ sequences randomly selected from the *A. thaliana* genome contains a true miRNA. The most difficult data set consisted of *A. thaliana* mRNAs; with these novoMIR reached a specificity of 0.975 (see Table 2). With all other data sets specificity was from 0.98 up to 1.00.

4.4. A Search for Pre-miRNAs in the Genome of Arabidopsis Thaliana. We wanted to test the program with a more realistic scenario, given the satisfying sensitivity and specificity values of novoMIR with our test data (see Tables 1 and 2). We selected all intergenic and intronic regions of the *A. thaliana* genome from "The Arabidopsis Information Resource" (TAIR), removed all pre-miRNA sequences, and searched within the remaining sequences for potential pre-miRNAs *via* novoMIR. novoMIR classified 828 sequences from the 30,413 intergenic sequences and 649 sequences from the 148,558 intronic sequences, respectively, as potential pre-miRNAs.

Despite this pleasingly low numbers of hits, however, an interpretation of this outcome is not easy. To get an impression on the hits, we searched with these potential pre-miRNA sequences with BLAST for any annotation and for the miRNA-typical expression pattern in the "*Arabidopsis* Small RNA Project Database" (ASRP) [28, 29]; such a typical expression pattern of a pre-miRNA includes sequences for the miRNA as well as for the miRNA* (for an example see supplemental Figure S7). To our surprise, we detected that some of the predicted candidates are already described as true pre-miRNAs. An example of such a sequence, predicted by novoMIR as a potential pre-miRNA, is located on *A. thaliana* chromosome 3 in the region between genes At3G09280 and At3G09290. Its secondary structure and its support by expressed small RNAs are shown in Figure 3 and Figure S8, respectively. It is already known as pre-miR2111a [30, 31], but not present in miRBase 14. The sequences of the mature miR2111a and of miR2111a* predicted by novoMIR also coincide with the sequences given in [30].

In the following, we mention shortly three further candidate hits, for which we found some support by small-RNA expression in the ASRP but no explicit annotation. One novoMIR hit is located on chromosome 4 between At4G22760 and At4G22770 close to the 3′ terminus of the latter, but on the opposite strand; for further details, see Figure 3 and Figure S9. The next hit (see Figure 3 and Figure S10) is located in between At5G52689 and At5G52690. The last mentioned hit is located in an intron of AT1G01650, which encodes for an aspartic-type endopeptidase/peptidase; the structure of this sequence is shown in Figure 3 and the expression pattern of the genomic region in Figure S11.

Several candidate hits have no support by small RNAs in the ASRP. It is known that many miRNAs are induced by biotic and abiotic stress [36–38]. Thus, a lack of small RNAs might either point to a false-positive prediction or to a stress condition not analyzed for expression of small RNAs. Further candidate hits are located in regions showing expression patterns similar to those of repetitive elements. A recently published review [39] discussed the possibility that some miRNAs could be evolved from repetitive genomic elements and/or duplication of genomic regions.

4.5. Viroids as Pre-miRNAs?

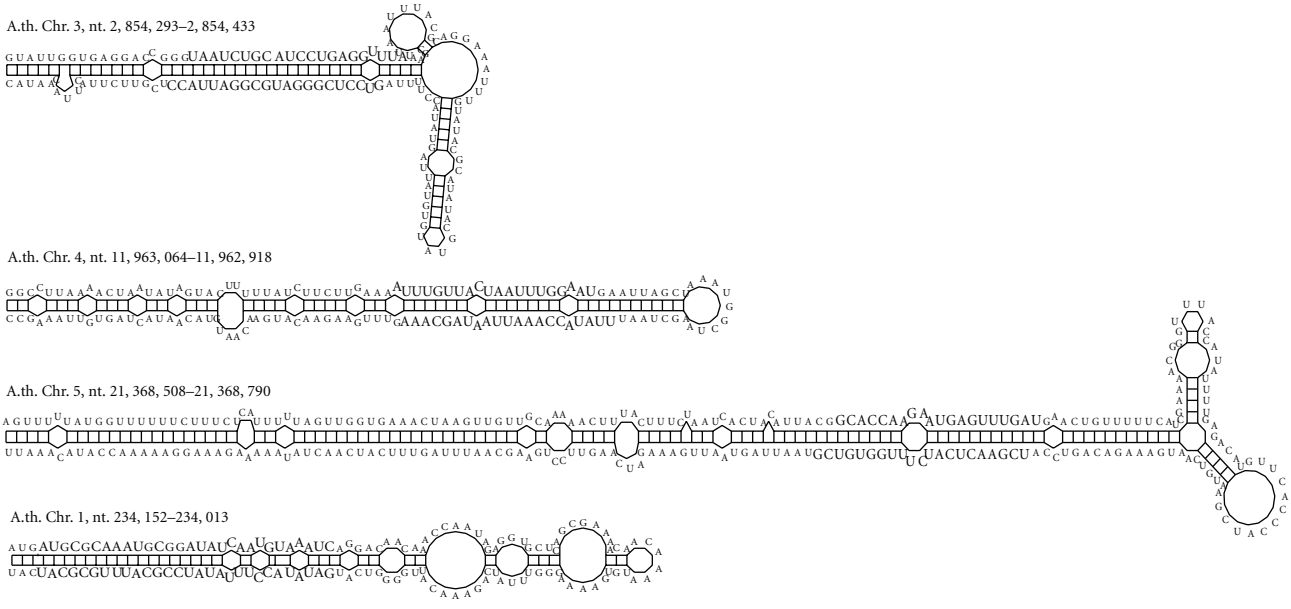
Viroids are plant-infectious, noncoding, unencapsidated, circular RNAs that are transcribed in a rolling-circle mechanism either in nuclei (*Pospiviroidae*) or in chloroplasts (*Avsunviroidae*) of infected plants. Viroids cause the production of viroid-specific small RNAs (vsRNA) similar in size to small interfering (siRNA) and miRNAs, but they do escape the cytoplasmic silencing mechanism. A positive (or negative) novoMIR prediction of viroids as potential pre-miRNAs would point to the genesis of vsRNAs. For further details, see recent reviews [40–43].

Potato spindle tuber viroid (PSTVd) is the type strain of *Pospiviroidae*. Because of its high self-complementarity the circular PSTVd RNA folds into a rod-like secondary structure of high thermodynamic stability (see Figure 4). This structure can be divided into five structural domains on the basis of homology between different pospiviroids [34]. Most sequence variants or strains of PSTVd differ by mutations in the pathogenicity-modulating (P) domain and/or variable (V) domain. Only a few nucleotide changes in the P domain are sufficient to exhibit remarkably different symptoms in infected tomato plants *Solanum lycopersicon* cv Rutgers. If this P domain would be the source of miRNA-like

Table 2: Specificity of novoMIR.

Data set	# sequences	Specificity[5]
A. thaliana mRNAs	710	0.975
noncoding RNAs[1]	1,296	1.000
noncoding RNAs[2]	455	0.982
shuffled *A. thaliana* pre-miRNAs	2,760	0.998
A. thaliana repetitive elements[3]	56	0.983
H. sapiens pseudohairpins	8,000	0.990
A. thaliana pseudohairpins	10,000	0.991
A. thaliana[4]	50,000	1.000

[1] 631 *A. thaliana* tRNAs, 63 5.8 S rRNAs, 602 5 S rRNAs
[2] noncoding RNAs from Rfam
[3] in total 134,000 nt
[4] $10 \times 5,000$ sequences of a length between 80 and 800 nt s randomly selected from the five chromosomes of *A. thaliana*
[5] Specificity is calculated as $TN/(TN + FP)$.



Chr.	$\Delta G^0_{\text{mfe},37°\text{C}}$ (kcal/mol)	L (nt)	f_{GC}	$\Delta G/L/f_{\text{GC}}$ (kcal/mol/nt)	Filters WD (bp)	HP (states)	Bp (bp)	R
3	−62.5	141	0.4	−1.12	0.96	44	20	3.2
4	−49.6	147	0.27	−1.24	0.91	70	11	2.1
5	−125.12	283	0.32	−1.38	0.94	118	22	2.4
1	−52.44	140	0.39	−0.97	0.91	68	21	2.1

Figure 3: Secondary structure and features of sequences classified by novoMIR as pre-miRNAs. Positions in the *A. thaliana* chromosomes is given above the structures. The predicted miRNA/miRNA* complexes are shown in larger italic characters. The table contains features of the sequences and their structures; the filter values are fraction of base pairs in five windows (with default threshold $t_{\text{WD}} = 0.65$), stem-loop length ($t_{\text{HP}} = 30$), maximal helix length ($t_{\text{BP}} = 8$), ratio of sequence and stem length ($t_{\text{R}} = 6$). For expression pattern of small RNAs at the genomic location of these sequences, see supplemental Figures S8–S11.

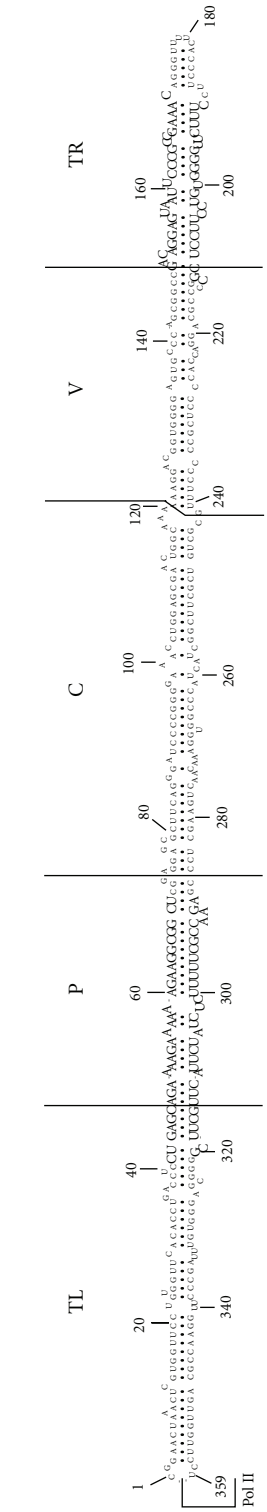


Figure 4: Secondary structure of PSTVd and location of miRNA/miRNA* complexes as predicted by novoMIR. The structure scheme is based on a consensus structure of 45 (+)-stranded circular PSTVd sequence variants [32]; the sequence is given for the PSTVd variant Intermediate [33]. The five homology domains of pospiviroids are marked as proposed by Keese and Symons [34]: terminal left and right (TL,TR), pathogenicity-modulating (P), central conserved (C), and variable (V) domain. The transcription start site for (−)-strand synthesis is marked by "polII" [35]. The predicted miRNA/miRNA* complexes are shown as larger italic characters. The complex in the P domain is predicted using default parameters; the complex in the TR domain is additionally predicted with a normalized energy threshold $t_{\Delta G/L/f_{GC}} = 0.69$ (instead of 0.75).

vsRNAs, these could interfere somehow with the host's metabolism leading to symptom production.

For an RNA with PSTVd sequence from positions 263–359/1–96, which is one of the structural elements present during processing of (+)-strand replication intermediates to circles [44], NOVOMIR predicted miRNA/miRNA* complexes in the P domain of PSTVd; for an RNA from positions 103–255, which is also a structural elements during processing, NOVOMIR predicted a further miRNA/miRNA* complex in the TR domain, but only after lowering the normalized energy threshold from the default value $t_{\Delta G/L/f_{GC}} = 0.75$ to 0.69. Both regions are marked by italic characters in Figure 4. NOVOMIR predicted identical positions for complexes in a full-length, linear PSTVd (1–359). Especially the prediction of vsRNAs derived from the P domain supports an involvement of vsRNAs in symptom production *via* vsRNA-induced (mis)regulation of plant-endogenous RNAs like mRNAs coding for transcription factors. This hypothesis is supported by deep-sequencing of PSTVd-derived vsRNAs in PSTVd-infected tomato plants (Diermann, Matoušek, Teune, Riesner and Steger, submitted) and sequencing of vsRNAs produced *in vitro* by DCL processing of PSTVd [45] which showed clusters of vsRNAs derived from the P domain. In contrast, [45, 46] found only vsRNAs in PSTVd-infected tomato plants that clustered in regions outside of the P domain. This discrepancy is unresolved but might be based for example on different purification procedures of the vsRNAs.

5. Conclusion

Plant pre-miRNAs are more heterogeneous in size and structure than animal pre-miRNAs but still show sufficient characteristic features—such as relative thermodynamic stability of their structure, length of helices, and number and size of loops—to be differentiated from other RNAs. Based on several of these features, we developed a series of filter steps and a statistical model that together are able to detect pre-miRNAs with a sensitivity of about 0.8 and a specificity of about 0.99. Thus, the program, which we call NOVOMIR, is well suited to search on a genomic scale for new pre-miRNAs that are not necessarily evolutionarily conserved. As an example, we searched with NOVOMIR for pre-miRNAs in nontranslated regions of the *A. thaliana* genome and detected among the high-scoring sequences experimentally verified pre-miRNAs, which were not annotated in the recent version of MIRBASE. Additionally, NOVOMIR recognizes viroids as pre-miRNAs, which supports the hypothesis that viroid-specific small RNAs are generated in a miRNA-like pathway.

Funding

The project was supported by a grant from the German Science Foundation to Detlev Riesner and G. Steger.

Competing Interests

The authors declare that they have no competing interests.

Acknowledgments

Acknowledgment is given to Dr. M. Schmitz and Dr. L. Nagel for critical reading of the paper. The package and supplementary information can be downloaded at http://www.biophys.uni-duesseldorf.de/novomir/.

References

[1] V. Ramachandran and X. Chen, "Small RNA metabolism in *Arabidopsis*," *Trends in Plant Science*, vol. 13, no. 7, pp. 368–374, 2008.

[2] O. Voinnet, "Origin, biogenesis, and activity of plant microRNAs," *Cell*, vol. 136, no. 4, pp. 669–687, 2009.

[3] V. Ambros, "A hierarchy of regulatory genes controls a larva-to-adult developmental switch in *C. elegans*," *Cell*, vol. 57, no. 1, pp. 49–57, 1989.

[4] B. J. Reinhart, F. J. Slack, M. Basson et al., "The 21-nucleotide let-7 RNA regulates developmental timing in *Caenorhabditis elegans*," *Nature*, vol. 403, no. 6772, pp. 901–906, 2000.

[5] J. R. Brown and P. Sanseau, "A computational view of microRNAs and their targets," *Drug Discovery Today*, vol. 10, no. 8, pp. 595–601, 2005.

[6] N. D. Mendes, A. T. Freitas, and M.-F. Sagot, "Current tools for the identification of miRNA genes and their targets," *Nucleic Acids Research*, vol. 37, no. 8, pp. 2419–2433, 2009.

[7] M. Yousef, L. Showe, and M. Showe, "A study of microRNAs in silico and in vivo: bioinformatics approaches to microRNA discovery and target identification," *FEBS Journal*, vol. 276, no. 8, pp. 2150–2156, 2009.

[8] W. Ritchie, M. Legendre, and D. Gautheret, "RNA stem-loops: to be or not to be cleaved by RNAse III," *RNA*, vol. 13, no. 4, pp. 457–462, 2007.

[9] T. Dezulian, M. Remmert, J. F. Palatnik, D. Weigel, and D. H. Huson, "Identification of plant microRNA homologs," *Bioinformatics*, vol. 22, no. 3, pp. 359–360, 2006.

[10] S. Pfeffer, A. Sewer, M. Lagos-Quintana et al., "Identification of microRNAs of the herpesvirus family," *Nature Methods*, vol. 2, no. 4, pp. 269–276, 2005.

[11] X. J. Wang, J. L. Reyes, N. H. Chua, and T. Gaasterland, "Prediction and identification of Arabidopsis thaliana microRNAs and their mRNA targets," *Genome Biology*, vol. 5, no. 9, p. R65, 2004.

[12] M. W. Jones-Rhoades and D. P. Bartel, "Computational identification of plant microRNAs and their targets, including a stress-induced miRNA," *Molecular Cell*, vol. 14, no. 6, pp. 787–799, 2004.

[13] S. Kadri, V. Hinman, and P. V. Benos, "HHMMiR: efficient de novo prediction of microRNAs using hierarchical hidden Markov models," *BMC Bioinformatics*, vol. 10, no. 1, article S35, 2009.

[14] C. Xue, F. Li, T. He, G.-P. Liu, Y. Li, and X. Zhang, "Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine," *BMC Bioinformatics*, vol. 6, article 310, 2005.

[15] S. Griffiths-Jones, R. J. Grocock, S. van Dongen, A. Bateman, and A. J. Enright, "miRBase: microRNA sequences, targets and gene nomenclature," *Nucleic Acids Research*, vol. 34, pp. D140–D144, 2006.

[16] S. Griffiths-Jones, H. K. Saini, S. Van Dongen, and A. J. Enright, "miRBase: tools for microRNA genomics," *Nucleic Acids Research*, vol. 36, no. 1, pp. D154–D158, 2008.

[17] B. H. Zhang, X. P. Pan, S. B. Cox, G. P. Cobb, and T. A. Anderson, "Evidence that miRNAs are different from other RNAs," *Cellular and Molecular Life Sciences*, vol. 63, no. 2, pp. 246–254, 2006.

[18] I. L. Hofacker, "Vienna RNA secondary structure server," *Nucleic Acids Research*, vol. 31, no. 13, pp. 3429–3431, 2003.

[19] S. Griffiths-Jones, S. Moxon, M. Marshall, A. Khanna, S. R. Eddy, and A. Bateman, "Rfam: annotating non-coding RNAs in complete genomes," *Nucleic Acids Research*, vol. 33, pp. D121–D124, 2005.

[20] S. R. Eddy, "SQUID—C function library for sequence analysis," 2008, http://selab.janelia.org/software.html#squid.

[21] A. F. A. Smit, R. Hubley, and P. Green, "RepeatMasker Open-3.0," 2004, http://www.repeatmasker.org/.

[22] K. L. S. Ng and S. K. Mishra, "De novo SVM classification of precursor microRNAs from genomic pseudo hairpins using global and intrinsic folding measures," *Bioinformatics*, vol. 23, no. 11, pp. 1321–1330, 2007.

[23] M. Garcia-Hernandez, T. Z. Berardini, G. Chen et al., "TAIR: a resource for integrated Arabidopsis data," *Functional and Integrative Genomics*, vol. 2, no. 6, pp. 239–253, 2002.

[24] R. Giegerich, B. Voß, and M. Rehmsmeier, "Abstract shapes of RNA," *Nucleic Acids Research*, vol. 32, no. 16, pp. 4843–4851, 2004.

[25] P. Steffen, B. Voß, M. Rehmsmeier, J. Reeder, and R. Giegerich, "RNAshapes: an integrated RNA analysis package based on abstract shapes," *Bioinformatics*, vol. 22, no. 4, pp. 500–503, 2006.

[26] I. L. Hofacker, B. Priwitzer, and P. F. Stadler, "Prediction of locally stable RNA secondary structures for genome-wide surveys," *Bioinformatics*, vol. 20, no. 2, pp. 186–190, 2004.

[27] J.-W. Nam, K.-R. Shin, J. Han, Y. Lee, V. N. Kim, and B.-T. Zhang, "Human microRNA prediction through a probabilistic co-learning model of sequence and structure," *Nucleic Acids Research*, vol. 33, no. 11, pp. 3570–3581, 2005.

[28] A. M. Gustafson, E. Allen, S. Givan, D. Smith, J. C. Carrington, and K. D. Kasschau, "ASRP: the *Arabidopsis* Small RNA Project database," *Nucleic Acids Research*, vol. 33, pp. D637–D640, 2005.

[29] T. W. H. Backman, C. M. Sullivan, J. S. Cumbie et al., "Update of ASRP: the *Arabidopsis* Small RNA Project database," *Nucleic Acids Research*, vol. 36, no. 1, pp. D982–D985, 2008.

[30] B. D. Pant, M. Musialak-Lange, P. Nuc et al., "Identification of nutrient-responsive Arabidopsis and rapeseed microRNAs by comprehensive real-time polymerase chain reaction profiling and small RNA sequencing," *Plant Physiology*, vol. 150, no. 3, pp. 1541–1555, 2009.

[31] N. Fahlgren, C. M. Sullivan, K. D. Kasschau et al., "Computational and analytical framework for small RNA profiling by high-throughput sequencing," *RNA*, vol. 15, no. 5, pp. 992–1002, 2009.

[32] G. Steger and D. Riesner, "Properties of viroids: molecular characteristics," in *Viroids*, A. Hadidi, R. Flores, J. W. Randles, and J. S. Semancik, Eds., pp. 15–29, CSIRO Publishing, Melbourne, Australia, 2003.

[33] H. J. Gross, H. Domdey, C. Lossow, et al., "Nucleotide sequence and secondary structure of potato spindle tuber viroid," *Nature*, vol. 273, no. 5659, pp. 203–208, 1978.

[34] P. Keese and R. H. Symons, "Domains in viroids: evidence of intermolecular RNA rearrangements and their contribution to viroid evolution," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 82, no. 14, pp. 4582–4586, 1985.

[35] N. Kolonko, O. Bannach, K. Aschermann et al., "Transcription of potato spindle tuber viroid by RNA polymerase II starts in the left terminal loop," *Virology*, vol. 347, no. 2, pp. 392–404, 2006.

[36] B. Li, W. Yin, and X. Xia, "Identification of microRNAs and their targets from *Populus euphratica*," *Biochemical and Biophysical Research Communications*, vol. 388, no. 2, pp. 272–277, 2009.

[37] L. I. Shukla, V. Chinnusamy, and R. Sunkar, "The role of microRNAs and other endogenous small RNAs in plant stress responses," *Biochimica et Biophysica Acta*, vol. 1779, no. 11, pp. 743–748, 2008.

[38] G. Jagadeeswaran, A. Saini, and R. Sunkar, "Biotic and abiotic stress down-regulate miR398 expression in Arabidopsis," *Planta*, vol. 229, no. 4, pp. 1009–1014, 2009.

[39] M. J. Axtell and J. L. Bowman, "Evolution of plant microRNAs and their targets," *Trends in Plant Science*, vol. 13, no. 7, pp. 343–349, 2008.

[40] E. M. Tsagris, Á. E. M. de Alba, M. Gozmanova, and K. Kalantidis, "Viroids," *Cellular Microbiology*, vol. 10, no. 11, pp. 2168–2179, 2008.

[41] B. Ding and A. Itaya, "Viroid: a useful model for studying the basic principles of infection and RNA biology," *Molecular Plant-Microbe Interactions*, vol. 20, no. 1, pp. 7–20, 2007.

[42] M. Schmitz and G. Steger, "Potato spindle tuber viroid (PSTVd)," *Plant Viruses*, vol. 1, pp. 106–115, 2007.

[43] J.-A. Daròs, S. F. Elena, and R. Flores, "Viroids: an Ariadne's thread into the RNA labyrinth," *EMBO Reports*, vol. 7, no. 6, pp. 593–598, 2006.

[44] T. Baumstark, A. R. W. Schröder, and D. Riesner, "Viroid processing: switch from cleavage to ligation is driven by a change from a tetraloop to a loop E conformation," *EMBO Journal*, vol. 16, no. 3, pp. 599–610, 1997.

[45] A. Itaya, X. Zhong, R. Bundschuh et al., "A structured viroid RNA serves as a substrate for dicer-like cleavage to produce biologically active small RNAs but is resistant to RNA-induced silencing complex-mediated degradation," *Journal of Virology*, vol. 81, no. 6, pp. 2980–2994, 2007.

[46] F. Di Serio, A.-E. M. De Alba, B. Navarro, A. Gisel, and R. Flores, "RNA-dependent RNA polymerase 6 delays accumulation and precludes meristem invasion of a viroid that replicates in the nucleus," *Journal of Virology*, vol. 84, no. 5, pp. 2477–2489, 2010.