Computational Biology and Chemistry 32 (2008) 400–405

Contents lists available at ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/compbiolchem

Research Article

Plant microRNA detection using the circular code information

Ahmed Ahmed, Christian J. Michel*

Equipe de Bioinformatique Théorique, LSIIT (UMR CNRS-ULP 7005), Université Louis Pasteur de Strasbourg, Pôle API, Boulevard Sébastien Brant, 67400 Illkirch, France

ARTICLE INFO

Article history:
Received 12 May 2008
Received in revised form 1 July 2008
Accepted 1 July 2008

Keywords:
Circular code
Plant microRNAs
Site identification
Computer method

ABSTRACT

In this paper we present a new computational method leading to the identification of a new property in the plant microRNAs. This property which is based on the circular code information is then used to detect microRNAs in plants. The common C^3 circular code X is a set of 20 trinucleotides identified in the reading frames of both eukaryotic and prokaryotic genes allowing retrieval of any frame in genes, locally anywhere in the three frames (reading frame and its two shifted frames) and automatically with the same window length of 13 nucleotides in each frame. This code X is detected around the beginning of microRNAs. This method based only on the internal structure of genes, i.e. the circular code, allows sensible and precise microRNA site identification in precursor microRNAs with a sliding window of only 14 nucleotides.

1. Introduction

1.1. MicroRNAs

MicroRNAs (miRNAs) are single-stranded RNA molecules of about 22 nucleotides in length. They regulate expression of target genes in a sequence-specific manner. miRNA genes were first identified in 1993 by loss-of-function mutations that cause defects in developmental timing in the *Caenorhabditis elegans* (Lee et al., 1993). They are found later in a wide variety of organisms ranging from plants, worms, animals and humans. They constitute, according to some estimates, up to 1% of the total number of genes in animal genomes and they are typically found either in inter-genetic areas or within introns of target genes (Lim et al., 2003).

1.1.1. Biogenesis

A miRNA is initially transcribed from a primary precursor molecule (pri-miRNA) which can be several hundreds to thousands of nucleotides long. All pri-miRNAs contain at least one characteristic hairpin like structure of about 70 nucleotides (stem-loop structure) known as precursor miRNA (pre-miRNA) (Fig. 1). The pre-miRNA is processed in the cell nucleus and is then exported to the cytoplasm. The exporters of the pre-miRNA are unknown but two candidate factors are proposed: exportin-t and exportin-5 (Grosshans and Slack, 2002; Gwizdek et al., 2003). In the cytoplasm, the hairpin structure of the pre-miRNA is recognized and cleaved by a complex containing nuclease Dicer in order to generate a mature miRNA (Bernstein et al., 2001; Chendrimada et al., 2005; Forstemann et al., 2005). The mature miRNA gets assembled into a protein effector complex called miRNP (miRNA containing ribonucleos–protein particles) which shares a lot of similarities to the effector complex called RISC (RNA-induced silencing complex) (Sontheimer, 2005).

Once the miRNP is assembled, the miRNA guides the complex to its target mRNA by base-pairing. In plants, miRNAs bind to a single and generally perfectly complementary coding region of the target mRNA. In contrast, most animal miRNAs bind to multiple and partially complementary sites in the 3′-UTRs (Bartel and Bartel, 2003; Zeng et al., 2003). The fate of the target mRNA is decided by the extent of base-pairing to the miRNA. A miRNA will direct the destruction of the target mRNA if it has perfect or near-perfect complementarity to the target (Hutvagner and Zamore, 2002). On the other hand, the presence of multiple partially complementary sites in the target mRNA will lead to an inhibition of protein accumulation without strongly affecting the mRNA level (Bartel and Chen, 2004).

1.1.2. miRNA Identification

Today, more than 400 miRNAs have been experimentally identified in mammalian genomes whereas estimates go up to 1000 and beyond. They have been detected by cloning and sequencing, computational approaches or as putative miRNAs homologous to miRNAs in other species.

The first miRNA, *lin-4*, was discovered during a biological investigation. It is a gene known to control the timing of *C. elegans* larval development. It does not code a protein but instead produces a pair of small RNAs. The shorter RNA is approximately 22 nucleotides in

* Corresponding author. Tel.: +33 3 90 24 44 62.
E-mail addresses: ahmed@dpt-info.u-strasbg.fr (A. Ahmed), michel@dpt-info.u-strasbg.fr (C.J. Michel).

1476-9271/$ – see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compbiolchem.2008.07.004

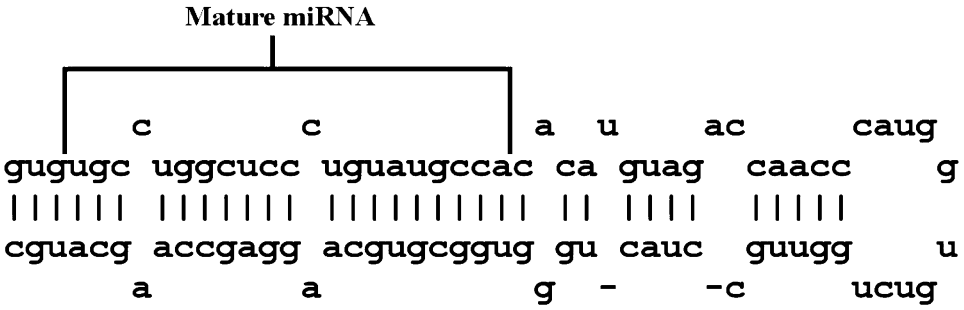

Fig. 1. An example of miRNA: the pre-miRNA of *Oryza sativa* miR160a (MI0000663) and its (mature) miRNA.

length and the other, about 61 nucleotides. The longer RNA was predicted to fold into a stem-loop structure proposed to be the precursor of the shorter one. The shorter RNA is now recognized as the founding member of an abundant class of tiny regulatory RNAs called miRNAs (Lee et al., 1993).

A particular cloning procedure developed in the *Drosophila melanogaster* genome (reverse transcription polymerase chain reaction amplification, concatemerization, cloning and sequencing) has identified 16 miRNAs (Lagos-Quintana et al., 2001).

Another approach to detect miRNAs has been proposed in human genome. It is based on the following steps: (i) computationally scanning the entire human genome for hairpin structures; (ii) annotating all hairpins for conserved, repetitive and protein coding regions; (iii) scoring hairpins by thermodynamic stability and structural features; (iv) determining the expression of computationally predicted miRNAs by a high-throughput miRNA microarray in several tissues (placenta, testis, thymus, brain and prostate); and (v) validating the sequences of predicted miRNAs giving high signals on microarrays using a new sequence-directed cloning and sequencing method. This approach has recognized 89 human miRNA genes (Bentwich et al., 2005).

Another research work has based their detection on: (i) testing of computationally predicted microRNAs by a modified microarray-based detection system; and (ii) cloning and sequencing large numbers of small RNAs from different human and mouse tissues. This method has led to 348 novel mouse and 81 novel human miRNA candidate genes (Berezikov et al., 2006).

The approach developed here is based on the assumption that the property of the near-perfect complementary of miRNAs with the protein coding region of its target mRNAs in plants will imply that the miRNAs have some structural properties of its target. Particularly, as the common C^3 circular code X has been statistically identified in genes of eukaryotes and prokaryotes, this property might be found in miRNAs. A similar and successful approach based on this code X allows the detection of frameshift sites in genes (Ahmed et al., 2007).

1.2. Common circular code

1.2.1. Identification

In 1996, a simple occurrence study of the 64 trinucleotides $\mathcal{T} = \{\text{AAA},\ldots,\text{TTT}\}$ in the three frames of genes has shown that the trinucleotides are not uniformly distributed in these three frames. We call frame 0 and frames 1 and 2, the reading frame of genes and its two shifted frames by 1 and 2 nucleotides in the 5′–3′ direction, respectively. By excluding the four trinucleotides with identical nucleotides $\mathcal{T}_{\text{id}} = \{\text{AAA,CCC,GGG,TTT}\}$ and by assigning each trinucleotide to a preferential frame (frame of its highest occurrence frequency), the same three subsets X_0, X_1, and X_2 of 20 trinucleotides are identified in the frames 0, 1 and 2, respectively, of two large and different gene populations of eukaryotes (26,757 sequences, 11,397,678 trinucleotides) and prokaryotes (13,686 sequences, 4,708,758 trinucleotides) (Arquès and Michel, 1996). These three trinucleotide subsets present several strong biomathematical properties, particularly the property of circular code. The subset $X_0 = \{\text{AAC,AAT,ACC,ATC,ATT,CAG,CTC,CTG,GAA,GAC,GAG,GAT,GCC,GGC,GGT,GTA,GTC,GTT,TAC,TTC}\}$ of 20 trinucleotides is the circular code identified in frame 0. Furthermore, it is a common circular code as it is observed in the reading frames of both eukaryotic and prokaryotic genes and in addition a C^3 code as its two permuted sets X_1 and X_2 are also circular codes (Table 1). X_0 is simply noted C^3 code X.

Table 1
The common C^3 circular code X identified in both eukaryotic and prokaryotic genes: X_0, X_1 and X_2 are the preferential sets of 20 trinucleotides in frames 0, 1 and 2 of genes, respectively

X_0	AAC AAT ACC ATC ATT CAG CTC CTG GAA GAC GAG GAT GCC GGC GGT GTA GTC GTT TAC TTC
X_1	ACA ATA CCA TCA TTA AGC TCC TGC AAG ACG AGG ATG CCG GCG GTG TAG TCG TTG ACT TCT
X_2	CAA TAA CAC CAT TAT GCA CCT GCT AGA CGA GGA TGA CGC CGG TGG AGT CGT TGT CTA CTT

These three trinucleotide subsets are obviously (law of large numbers) retrieved in these two gene populations with the actual statistical studies (results not shown).

1.2.2. Main Properties

We recall the main properties of this common C^3 code X which will be involved in this paper. A recent review of circular codes in genes details the research context, the history and their different properties (Michel, 2008).

Definition 1. The (left circular) permutation $\mathcal{P}$ of a trinucleotide $w_0 = l_0l_1l_2$, $w_0 \in \mathcal{T}$, is the permuted trinucleotide $\mathcal{P}(w_0) = w_1 = l_1l_2l_0$, e.g. $\mathcal{P}(\text{AAC}) = \text{ACA}$, and $\mathcal{P}(\mathcal{P}(w_0)) = w_2 = l_2l_0l_1$, e.g. $\mathcal{P}(\mathcal{P}(\text{AAC})) = \text{CAA}$. This definition is naturally extended to the trinucleotide set permutation: the permutation $\mathcal{P}$ of a set of trinucleotides is the permuted trinucleotide set obtained by the permutation $\mathcal{P}$ of all its trinucleotides.

Definition 2. The complementarity $\mathcal{C}$ of a trinucleotide $w_0 = l_0l_1l_2$, $w_0 \in \mathcal{T}$, is the complementary trinucleotide $\mathcal{C}(w_0) = \mathcal{C}(l_2)\mathcal{C}(l_1)\mathcal{C}(l_0)$ with $\mathcal{C}(\text{A}) = \text{T}$, $\mathcal{C}(\text{C}) = \text{G}$, $\mathcal{C}(\text{G}) = \text{C}$, $\mathcal{C}(\text{T}) = \text{A}$, e.g. $\mathcal{C}(\text{AAC}) = \text{GTT}$. This definition is also naturally extended to the trinucleotide set complementarity.

Notation 1. $\mathcal{A}$ being a finite alphabet, $\mathcal{A}^*$ denotes the words over $\mathcal{A}$ of finite length including the empty word ε of length 0 and $\mathcal{A}^+$, the words over $\mathcal{A}$ of finite length greater or equal to 1. Let x_1x_2 be the concatenation of the two words x_1 and x_2.

Definition 3. A subset X of $\mathcal{A}^+$ is a circular code if $\forall n$, $m \geq 1$ and $x_1, x_2, \ldots, x_n, y_1, y_2, \ldots, y_m \in X$ and $r \in \mathcal{A}^*$, $s \in \mathcal{A}^+$, the equalities $sx_2 \cdots x_n r = y_1y_2 \cdots y_m$ and $x_1 = rs$ imply $n = m$, $r = \varepsilon$ and $x_i = y_i$, $1 \leq i \leq n$.

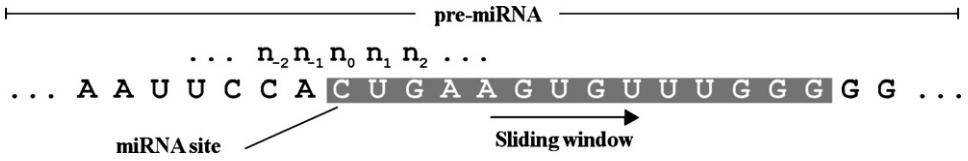

Fig. 2. The method reading frame is $\ldots n_{-1}n_0n_1\ldots$. The miRNA site begins at $i=0$ and is noted n_0.

A circular code allows the reading frames in genes to be retrieved. It is a set of words over an alphabet such that any word written on a circle (the next letter after the last letter of the word being the first letter) has a unique decomposition into words of the circular code. As an example, let the set Y be composed of the six following words: $Y=\{$AAT,ATG,CCT,CTA,GCC,GGC$\}$ and the word w, be a series of the nine following letters: $w=$ ATGGCCCTA. The word w, written on a circle, can be factorized into words of Y according to two different ways: ATG,GCC,CTA and AAT,GGC,CCT, the commas showing the way of decomposition. Therefore, Y is not a circular code. In contrast, if the set Z obtained by replacing the word GGC of Y by GTC is considered, i.e. $Z=\{$AAT,ATG,CCT,CTA,GCC,GTC$\}$, then there will never exist an ambiguity with any word of Z, particularly w is not ambiguous, and Z is a circular code. The construction frame of a word generated by any concatenation of words of a circular code can be retrieved after the reading, starting from anywhere in the generated word, of a certain number of nucleotides depending on the code. This series of nucleotides is called the window of the circular code. Then, the minimal window length is the size of the longest ambiguous word which can be read in at least two frames, more one letter.

Proposition 1. *Permutation*: $\mathcal{P}(X_0)=X_1$ *and* $\mathcal{P}(\mathcal{P}(X_0))=X_2$ (X_0 *generates* X_1 *by one permutation and* X_2 *by another permutation*).

Proposition 2. *Complementarity:* $\mathcal{C}(X_0)=X_0$ (X_0 *is self-complementary*) *and*, $\mathcal{C}(X_1)=X_2$ *and* $\mathcal{C}(X_2)=X_1$ (X_1 *and* X_2 *are complementary to each other*).

Proposition 3. C^3 *code:* if X_0, X_1 *and* X_2 *are circular codes then* X_0, X_1 *and* X_2 *are* C^3 *codes. As the code* X_0 *is coding the reading frame* (*frame* 0) *in genes, i.e. the most important frame, and as it is self-complementary* (Proposition 2), *it is considered as the main* C^3 *code and noted X simply*.

Remark 1. A circular code Y_0 does not necessarily imply that Y_1 and Y_2 obtained by permutation of Y_0 are also circular codes.

Proposition 4. *Rarity: the occurrence probability of the complementary* C^3 *code X is equal to* $216/3^{20}\approx 6\times 10^{-8}$. *Indeed, this probability is equal to the computed number of complementary* C^3 *codes* (216) *divided by the number of potential codes* ($3^{20}=3{,}486{,}784{,}401$).

Proposition 5. *Largest window length: the lengths of the minimal windows of* X_0, X_1 *and* X_2 *to retrieve the frames* 0, 1 *and* 2, *respectively, are all equal to* 13 *nucleotides and represent the largest window length among the* 216 *complementary* C^3 *codes*.

Proposition 6. *Common: the* C^3 *code X is identified in both eukaryotic and prokaryotic genes*.

Proposition 7. *As a consequence of the previous propositions, the common* C^3 *code X allows retrieval of any frame in genes, locally anywhere in the three frames and particularly without start codons in reading frames, and automatically with the same window length of* 13 *nucleotides in each frame*.

2. Method

2.1. Definition of a Statistical Function P Based on the Common C^3 Code X

The method presented here is based on the following concept: as the genes have a structural property of circular code and as the miRNAs and their target genes make a near-perfect complementary, then the miRNAs may also have some structural properties of their targets, particularly the C^3 circular code X.

We give here an algorithm that computes the occurrence of the C^3 code X on a sliding window that scans the entire pre-miRNA. Let $\mathcal{T}=\{$AAA,...,TTT$\}$ be the set of 64 trinucleotides over the 4-nucleotides alphabet $\mathcal{N}=\{$A,C,G,T$\}$. Let $n\in\mathcal{N}$ be a nucleotide, $t\in\mathcal{T}$, a trinucleotide and $\mathcal{F}$ a plant pre-miRNA population with $n(\mathcal{F})$ sequences s. Each sequence s contains a miRNA which starts at position $i=0$. Let n_0 be the first nucleotide of the miRNA in the sequence s which is also referred to as miRNA site (Fig. 2). By convention, all nucleotides before the miRNA site have a negative position $i<0$ and all nucleotides after the miRNA site, a positive position $i>0$. The sequence s is considered as a series of nucleotides n_i and the method reading frame is $\ldots n_{-1}n_0n_1\ldots$. Let $w_i=n_in_{i+1}\ldots n_{i+13}$ be a window of length $|w|=4$ trinucleotides plus two nucleotides added to compute the score function in the two other shifted frames, where n_i is the ith nucleotide in the sequence s. Let $t_i^{l,\phi}$ be the lth trinucleotide, $l\in\{0,1,2,3\}$, at frame ϕ, $\phi\in\{0,1,2\}$, in w_i (Fig. 3). The length of the sliding window w_i is the length of the minimal windows of the three circular codes X_0, X_1 and X_2 to retrieve the frames 0, 1 and 2 in genes, respectively (Proposition 5). Let X_f, $f\in\{0,1,2\}$, be the three codes X_0, X_1 and X_2 in the 3 frames f.

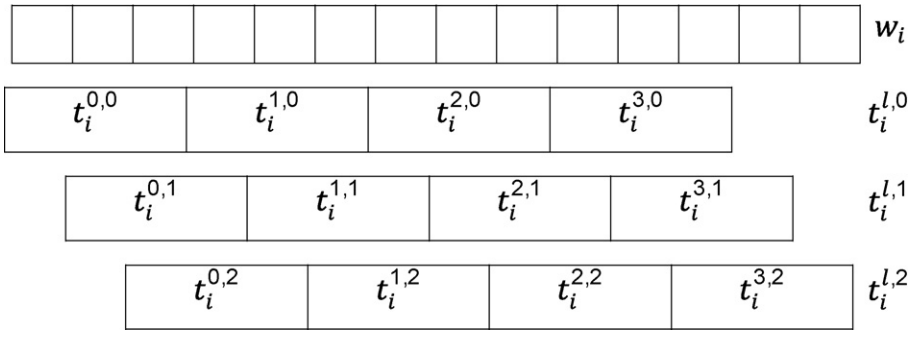

Fig. 3. $t_i^{l,\phi}$ is the lth trinucleotide, $l\in\{0,1,2,3\}$, at frame ϕ, $\phi\in\{0,1,2\}$, in a sliding window w_i.

In a given window w_i, the function $\delta_f(t_i^{l,\phi})$ indicates whether or not the trinucleotide $t_i^{l,\phi}$ belongs to the code X_f

$$\delta_f\left(t_i^{l,\phi}\right) = \begin{cases} 1 & \text{if } t_i^{l,\phi} \in X_f \\ 0 & \text{otherwise} \end{cases}.$$

Then, the score function $P(X_f, i, s)$ computes the occurrence of the code X_f in its associated frame, i.e. $\phi = f$, in a window w_i of a sequence s

$$P(X_f, i,\ s) = \frac{1}{|w|}\sum_{l=0}^{|w|-1}\delta_f(t_i^{l,f}).$$

The score function $P(i, s)$ computes the average occurrence of the C^3 code X in a window w_i of a sequence s

$$P(i,\ s) = \frac{1}{3}\sum_{f=0}^{2}P(X_f, i, s).$$

Finally, the score function $P(i, \mathcal{F})$ computes the occurrence of the C^3 code X in a window w_i for all sequences s of a population $\mathcal{F}$

$$P(i, \mathcal{F}) = \frac{1}{n(\mathcal{F})}\sum_{s\in\mathcal{F}}P(i, s).$$

Proposition 8. *If in each sequence s of $\mathcal{F}$ the trinucleotides $t_i^{l,\phi}$ of a window w_i belong to the set X_f such that $\phi = f$ then $P(i, \mathcal{F}) = 1$ and the occurrence of the C^3 code X is maximum.*

Proposition 9. *If in each sequence s of $\mathcal{F}$ the trinucleotides $t_i^{l,\phi}$ of a window w_i belong to the set X_f such that $\phi \neq f$ then $P(i, \mathcal{F}) = 0$ and the C^3 code X does not occur.*

Proposition 10. $0 \leq P(i, \mathcal{F}) \leq 1$ (*consequence of* Propositions 8 and 9).

2.2. Data Acquisition

The pre-miRNA sequences studied in our method are extracted from the miRNA registry release 11.0 released on April 2008 (Griffiths-Jones et al., 2006). This release contains 6396 entries. The plant population $\mathcal{F}$ studied in this release has 1466 pre-miRNA sequences (206 kb). The entries belong to various plant families including *Arabidopsis thaliana*, *Oryza sativa*, *Populus trichocarpa*, *Zea mays* and other families.

The lengths of pre-miRNAs and their miRNA sites vary with each sequence. This data variation implies that the developed method will be computed for an interval of nucleotides around the miRNA site $i = 0$ where there is a sufficient number of sequences (not detailed).

3. Results

3.1. Identification of a Periodicity Modulo 3 and a Circular Code Rich Region in Plant miRNAs

Fig. 4 shows the results of the function $P(i, \mathcal{F})$ applied to the plant miRNAs. The *X*-axis represents the position i of the sliding window w_i and the *Y*-axis, its value $P(i, \mathcal{F})$. A periodicity modulo 3 is observed from the nucleotide n_{-19} to n_{11} with a peak at the nucleotide n_{-1}. When the trinucleotides at frames 0, 1 and 2 of w_i belong to the sets X_0, X_1 and X_2, respectively, then the value $P(i, \mathcal{F})$ is high (Proposition 8) and obviously the values $P(i+1, \mathcal{F})$ and $P(i+2, \mathcal{F})$ for w_{i+1} and w_{i+2}, respectively, are low since the words of each set X_0, X_1 and X_2 are in a shifted frame (Proposition 9). The periodicity modulo 3 and the Propositions 8 and 9 imply that the C^3 code X occurs significantly around the miRNA sites.

In order to focus on the regions which conserve the property of circular code, a statistical function $Q(i, \mathcal{F})$ is defined showing the average of three consecutive differences of values $P(i, \mathcal{F})$

$$Q(i, \mathcal{F}) = \frac{1}{3}\sum_{j=i-1}^{i+1}|P(j, \mathcal{F}) - P(j-1, \mathcal{F})|.$$

Fig. 5 shows a circular code rich region in a wide interval of 20 nucleotides $i \in \{-11, \ldots, 8\}$ which also has a peak at $i = 0$ with a value $Q(0, \mathcal{F}) = 0.1177$.

3.2. Use of the Property of Circular Code to Detect Plant miRNAs

The application of the statistical function $Q(i, \mathcal{F})$ on individual sequences of the plant population $\mathcal{F}$ leads to the recognition of 332 plant miRNAs among the 1466 ones, i.e. about 23%, by using only this computational method (see also Section 4). As an example of

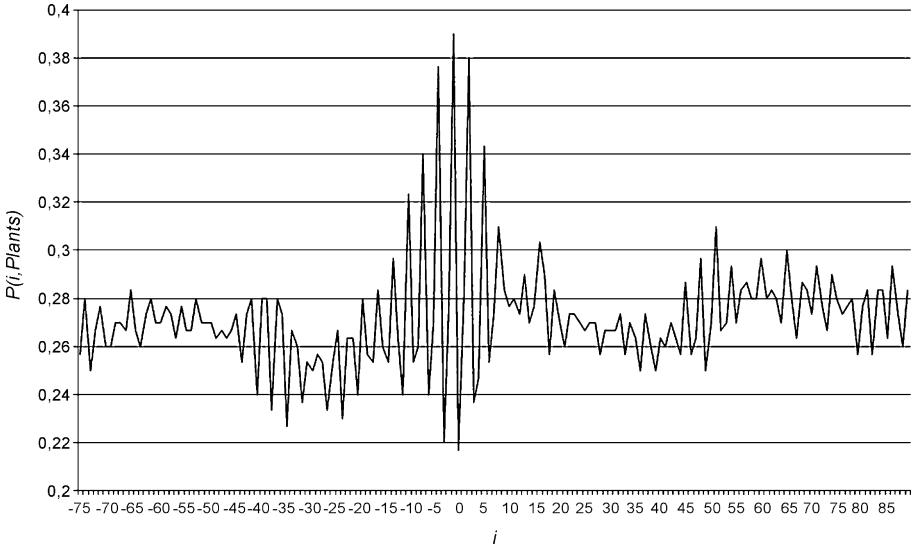

Fig. 4. Function $P(i, \mathcal{F})$ applied to the plant miRNAs $\mathcal{F}$. The *X*-axis represents the position i of the sliding window w_i and the *Y*-axis, its value $P(i, \mathcal{F})$. A periodicity modulo 3 starting from the nucleotide n_{-19} to n_{11} with a peak at the nucleotide n_{-1} is obviously detected.

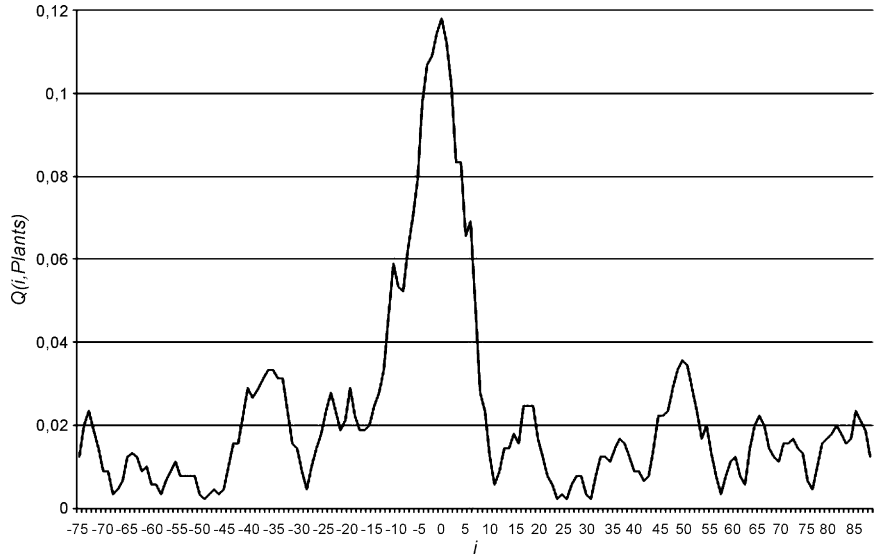

Fig. 5. Function $Q(i, \mathcal{F})$, average of three consecutive differences of values $P(i, \mathcal{F})$ (Fig. 4), applied to the plant miRNAs $\mathcal{F}$. The *X*-axis represents the position i of the sliding window w_i and the *Y*-axis, its value $Q(i, \mathcal{F})$. A circular code rich region is significantly identified in the nucleotide interval $i \in \{-11, \ldots, 8\}$ with a peak at $i = 0$ and a value $Q(0, \mathcal{F}) = 0.1177$.

Table 2
The *Medicago truncatula* miR395b pre-miRNA and its miRNA "MI0001747" starting at the nucleotide position 85 of its pre-miRNA (underlined)

Pre-miRNA	UACUUGUUGAUUUUCUCUUGGAGUUCCUCUGAAUGCUUCAAA-CAUGAGACAAUCUUGAUAGAAAUUAUGGAUAGUUCUUGUUCA-AUGAAGUAUUUGGGGGAACUCUUGGAAUUGAAUCAACAUA
miRNA	AUGAAGUAUUUGGGGGAACUC

a detected miRNA (at the sequence level), we present the results for the miRNA "MI0001747" ($\mathcal{F}$ ="*Medicago truncatula* miR395b"). Table 2 shows its complete pre-miRNA and miRNA sequences, the miRNA starting at the nucleotide position 85 from the beginning of its pre-miRNA.

Fig. 6 shows the function $Q(i, M.\ truncatula)$. A peak is identified at the position $i = -3$ with a value $Q(-3, M.\ truncatula) = 0.5$ in a circular code rich region in the nucleotide interval $i \in \{-10, \ldots, 0\}$.

4. Discussion

In this paper, a new computer method that searches for circular code rich regions has been developed. This method is based on the common C^3 code X identified in genes of eukaryotes and prokaryotes. It is very sensitive since it uses a sliding window of small length with four trinucleotides which is the length of the minimal windows of the three circular codes X_0, X_1 and X_2 to retrieve the frames 0, 1 and 2 in genes, respectively.

The application of this method on a population of plant pre-miRNAs identifies two new properties: a periodicity modulo 3 and a circular code rich region around their miRNA sites (in the nucleotide

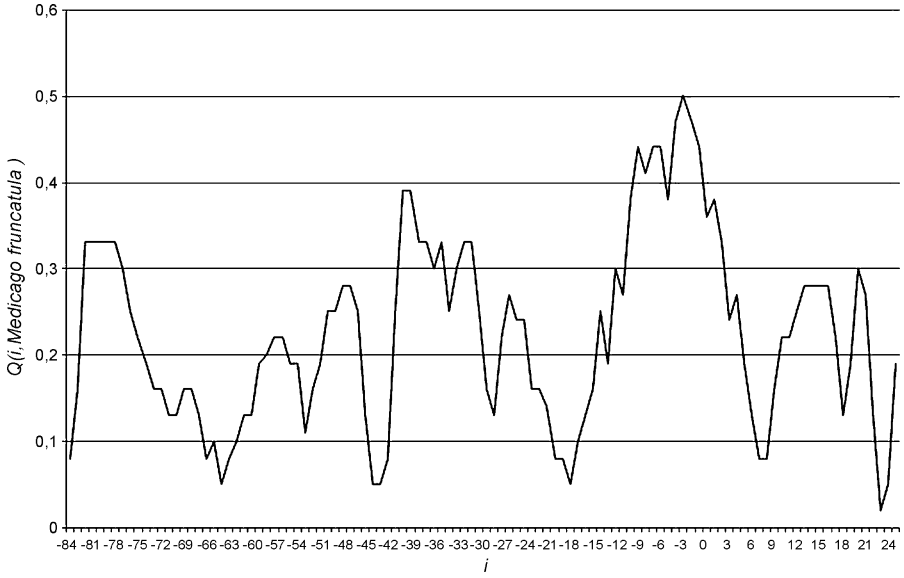

Fig. 6. Function $Q(i, \mathcal{F})$ on an individual sequence. Example with *Medicago truncatula*. The *X*-axis represents the position i of the sliding window w_i and the *Y*-axis, its value $Q(i, M.\ truncatula)$. A circular code rich region occurs in the nucleotide interval $i \in \{-10, \ldots, 0\}$ with a peak at $i = -3$ and a value $Q(-3, M.\ truncatula) = 0.5$.

interval $\{-11, \ldots, 8\}$). Thus, the common C^3 code X found in plant genes also occurs in their miRNAs.

The use of this identified property of circular code on individual sequences recognizes directly 23% of plant miRNAs. The remaining plant miRNAs may have non-maximal codes X, i.e. codes with subsets of trinucleotides (less than 20), or codes X with mutations or codes X with an incomplete complementarity property, etc. Such degenerated properties of the code X have not been yet investigated. Indeed, the aim of this paper mainly deals with the identification of this new property of circular code in plant miRNAs.

This method applied to a population of animal and human pre-miRNAs does not carry out the same results (results not shown). The lack of circular code rich regions in these pre-miRNAs may be due to the biological nature and structure of their miRNAs. Indeed, they bind to multiple, partially complementary sites in the 3′-UTRs of its target genes which are regions that are not characterized by the circular code property (Michel, 2008).

The developed method showing a circular code signal around the plant miRNA sites can be improved by varying the size of the sliding window, by using circular code information specific to the plant genomes, by considering variant codes X, etc. Its principle can also be associated to the existing methods to improve the detection of miRNAs.

References

Ahmed, A., Frey, G., Michel, C.J., 2007. Frameshift signals in genes associated with the circular code. Silico Biol. 7, 151–154.

Arquès, D.G., Michel, C.J., 1996. A complementary circular code in the protein coding genes. J. Theor. Biol. 182, 45–58.

Bartel, B., Bartel, D.P., 2003. MicroRNAs: at the root of plant development? Plant Physiol. 132, 709–717.

Bartel, D.P., Chen, C.Z., 2004. Micromanagers of gene expression: the potentially widespread influence of metazoan microRNAs. Nat. Rev. Genet. 5, 396–400.

Bentwich, I., Avniel, A., Karov, Y., Aharonov, R., Gilad, S., Barad, O., et al., 2005. Identification of hundreds of conserved and nonconserved human microRNAs. Nat. Genet. 37, 766–770.

Berezikov, E., van Tetering, G., Verheul, M., van de Belt, J., van Laake, L., Vos, J., Verloop, R., van de Wetering, M., Guryev, V., Takada, S., van Zonneveld, A.J., Mano, H., Plasterk, R., Cuppen, E., 2006. Many novel mammalian microRNA candidates identified by extensive cloning and RAKE analysis. Genome Res. 16, 1289–1298.

Bernstein, E., Caudy, A.A., Hammond, S.M., Hannon, G.J., 2001. Role for a bidentate ribonuclease in the initiation step of RNA interference. Nature 409, 363–366.

Chendrimada, T.P., Gregory, R.I., Kumaraswamy, E., Norman, J., Cooch, N., Nishikura, K., Shiekhattar, R., 2005. TRBP recruits the Dicer complex to Ago2 for microRNA processing and gene silencing. Nature 436, 740–744.

Forstemann, K., Tomari, Y., Du, T., Vagin, V.V., Denli, A.M., Bratu, D.P., Klattenhoff, C., Theurkauf, W.E., Zamore, P.D., 2005. Normal microRNA maturation and germline stem cell maintenance requires Loquacious, a double-stranded RNA-binding domain protein. PLoS Biol. 3, 1187–1201.

Griffiths-Jones, S., Grocock, R.J., van Dongen, S., Bateman, A., Enright, A.J., 2006. miRBase: microRNA sequences, targets and gene nomenclature. NAR 34 (Database Issue), D140–D144.

Grosshans, H., Slack, F., 2002. Micro-RNAs: small is plentiful. J. Cell Biol. 156, 17–21.

Gwizdek, C., Ossareh-Nazari, B., Brownawell, A.M., Doglio, A., Bertrand, E., Macara, I.G., Dargemont, C., 2003. Exportin-5 mediates nuclear export of minihelix-containing RNAs. J. Biol. Chem. 278, 5505–5508.

Hutvagner, G., Zamore, P.D., 2002. A microRNA in a multiple-turnover RNAi enzyme complex. Science 297, 2056–2060.

Lagos-Quintana, M., Rauhut, R., Lendeckel, W., Tuschl, T., 2001. Identification of novel genes coding for small expressed RNAs. Science 294, 853–858.

Lee, R.C., Feinbaum, R.L., Ambros, V., 1993. The *C. elegans* heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell 75, 843–854.

Lim, L.P., Glasner, M.E., Yekta, S., Burge, C.B., Bartel, D.P., 2003. Vertebrate microRNA genes. Science 299, 1540.

Michel, C.J., 2008. A review of circular codes in genes. Comput. Math. Appl. 55, 984–988.

Sontheimer, E.J., 2005. Assembly and function of RNA silencing complexes. Nat. Rev. Mol. Cell. Biol. 6, 127–138.

Zeng, Y., Yi, R., Cullen, B.R., 2003. miRNAs and small interfering RNAs can inhibit mRNA expression by similar mechanisms. Proc. Nat. Acad. Sci. U.S.A. 100, 9779–9784.