Full Length Research Paper

In silico identification of MicroRNAs and their targets in diatoms

Yong Zhong Lu* and Jie Liu

Biology Department, Qingdao University of Science and Technology, Qingdao, 266043, China.

As one of the most successful groups of microalgae in the contemporary ocean, diatoms are of ecological, biotechnological and evolutionary significance. Recent research has revealed for the first time the presence of functional silencing machinery in diatoms. Nevertheless, no microRNAs (miRNAs) participating in their gene regulation have been reported. Based on the principle of sequence conservation, previous known plant miRNAs were blasted against the expressed sequence tag databases of the marine diatoms *Phaeodactylum tricornutum* and *Thalassiosira pseudonana* and according to a series of filtering criteria, 6 conserved miRNAs were identified and 5 potential target genes of them were subsequently predicted. Alignmental and phylogenetic analyses showed that the miRNA precursor sequences were unexpectedly poorly conserved and distantly related to other family members. Above all, the findings from this study will contribute not only to further research of miRNAs features and regulatory mechanisms in diatoms, but also to their evolutionary research by virtue of new molecular tools.

Key words: Diatom, microRNA, evolution, bioinformatics.

INTRODUCTION

Diatoms are unicellular photoautotrophic eukaryotes that play an important role in ecology by fixing large amounts of carbondioxide (CO_2) and generate most of the organic matter that serves as food for life in the sea. They greatly influence global climate, atmospheric carbon dioxide concentration and marine ecosystem function (Armbrust, 2009). After about 100 million years of evolution, they have become one of the most successful groups of microalgae in the contemporary ocean and have a rather unusual cell biology and genetic constitution (Kroth, 2007; Vardi et al., 2008). They are also of biotech-nological interest in carbon neutral synthesis of fuels, pharmaceuticals, health food, biomolecules and materials relevant to nanotechnology and so on (Bozarth et al., 2009).

Considering the ecological and biotechnological signifi-cance of diatoms as well as their potential in helping to dissect the evolution of eukaryotes, efforts have been made in many aspects to reveal the ecology, evolution and metabolism secrets (Dagan et al., 2009; Maheswari

et al., 2009; Nunn et al., 2009). And due to the lack of tools for the study of diatomaceous gene function, successful silencing in *Phaeodactylum tricornutum* by expressing of antisense RNAs revealed for the first time the presence of a functional silencing machinery in diatoms and represented a major advance for under-standing their biology and ecology (Riso et al., 2009). However, the molecular basis for this machinery in diatoms is not largely unknown, nor have regulatory RNAs been reported.

MicroRNAs (miRNAs) are a class of endogenous, small, non-coding, single-stranded RNAs that act as post-transcriptional regulators in eukaryotes (Unver et al., 2009). They have been reported to be located mostly within noncoding regions of genomes and transcribed by RNA polymerase II (Baskerville et al., 2005; Reinhart et al., 2002). The generation of mature miRNA is a complicated enzyme-catalyzed process, from the initial transcript (pri-miRNA) to the precursor (pre-miRNA) with a characteristic hairpin structure, then a miRNA duplex (miRNA:miRNA*) (Bartel, 2004). In the end, it is assembled to the RNA-induced silencing complex (RISC) to direct its activity on a target mRNA, depending on the degree of base-pairing between the miRNA and the responsive element and results in either cleavage or

---

*Corresponding author. E-mail: luyz@yahoo.com. Tel: +86-532-84022845.

translational repression of the target mRNA. Perfect complementarity generally results in cleavage, such as in plants, whereas imperfect base-pairing leads to translational repression (Bartel, 2004; Carrington et al., 2003).

MiRNA genes represent about 1 - 2% of the known eukaryotic genomes and constitute an important class of fine-tuning regulators that are involved in several physiological or disease-associated cellular processes (Barbato et al., 2009). In plants, miRNAs are involved in a number of biological mechanisms including growth, development and defense response against every sort of stress (Rodriguez et al., 2010; Wang et al., 2009; Mathieu et al., 2009; Ding et al., 2009; Fornara et al., 2009; Pothof et al., 2009; Ruiz-Ferrer et al., 2009; Sunkar et al., 2007; Ambros et al., 2007; Lu et al., 2005; Chen, 2004). Considering the importance of miRNAs in gene regulation, two major categories of approaches have been applied for miRNAs investigation (Unver et al., 2009).

Compared to the experimental approaches, computation (bioinformatics) methods have been proved to be faster, more affordable and more effective, contributing mostly to today's plentiful storage in miRBase (Unver et al., 2009). Different computational miRNA finding strategies have been developed based on a core principle of looking for conserved sequences between different species that can fold into extended hairpins (Bonnet et al., 2004). The biogenesis of miRNAs suggests that it is possible to find miRNAs by searching expressed sequence tags (ESTs) with known miRNAs. There have been more and more reports about the identification of miRNAs by mining the repository of available ESTs (Han et al., 2009; Zhang et al., 2009; Song et al., 2009; He et al., 2008; Sunkar et al., 2008; Yin et al., 2008; Xie et al., 2007; Zhang et al., 2005). Furthermore, the pervasive and crucial roles of miRNAs have drawn more and more attentation in evolutionary researches (Kosik, 2009; Axtell et al., 2008; Zhang et al., 2006). Till now, except for the green algae *Chlamydomonas reinhardtii*, no miRNAs have been found in other algae. In this study, based on sequence conservation of known plant miRNAs, EST databases of *P. tricornutum* and *Thalassiosira pseudonana* were mined to find miRNAs. This work will not only help to know more about the roles of miRNAs in diatoms, but also provide potential molecular tools for gene function and evolution analyses.

MATERIALS AND METHODS

Sequences and software

The known plant miRNA sequences from *Arabidopsis*, *Brassica*, *Glycine*, *Saccharum*, *Sorghum*, *Vitis*, *Solanum*, *Oryza*, *Triticum*, *Chlamydomonas* and other plant species were downloaded from the miRNA database miRBase (http://www.mirbase.org) (Release 14th September, 2009). After removal of the repeated sequences, 2177 items were left as the reference set. The 133891 EST sequences of *P. tricornutum* and 61920 EST sequences of *T.*

pseudonana were downloaded from GenBank (www.ncbi.nlm.nih.gov/). Blast- 2.2.21-ia3 was downloaded from NCBI and set up locally, RNA secondary structure and the free energy was calculated by mfold online (http://mfold.bioinfo.rpi.edu/) (Zuker, 2003). The software MiRNA assist was applied to improve the analysis efficiency (Xie et al., 2007).

Computational prediction of diatom miRNAs

The prediction procedure was shown in Figure 1. The sequences of known plant miRNAs were used as query sequences for BLAST search against the EST databases, with the BLASTN parameters value being 1000 and word-match size between the query and data base sequences being 7. Mature miRNA sequences should be no less than 16 nt, and the mismatches should be less than 4. Wherever available, precursor sequences of 400 nt were extracted (200 nt upstream and 200 nt downstream to the BLAST hits) and used for the hairpin structure prediction. If the length of a sequence was less than 400 nt, the entire available sequence was used as a miRNA precursor sequence. These precursor sequences were then blasted online to remove the protein coding sequences (http://blast.ncbi.nlm.nih.gov/Blast.cgi). The retained precursor sequences under went hairpin structure prediction through mfold online. Only those meeting the following criteria were designated as miRNA homologs: (1) the RNA sequence folding into an appropriate stem-loop hairpin secondary structure, (2) a mature miRNA sequence located in one arm of the hairpin structure, (3) predicted mature miRNAs with no more than 3 nt substitutions as compared with the known miRNAs, (4) miRNAs having less than 6 mismatches with the opposite miRNA* sequence in the other arm, (5) no loop or break in miRNA* sequences, (5) predicted secondary structures with higher minimal folding free energy (MFE) and minimal folding free energy index (MFEI), the MEFI usually being about 0.85 according to the report (Zhang et al., 2006). Also, the AU content of pre-miRNA within the scope of 30 and 70% was considered (Xie et al., 2007).

Computational prediction of miRNA targets

MiRNA targets prediction was performed by aligning the predicted miRNA sequences with the diatom mRNA sequences through the BLAST program. The targets were screened according to these criteria: the number of mismatches should be less than 4, the minimal free energy of the pairing between miRNA and its target mRNA was lower than −28.2 kcal/mol (Ambros et al., 2003). After removal of the repeated sequences, the potential target genes were BLASTXed against protein databases online to predict their function (Identity > 25%), since there were no functional annotations for mRNA sequences of *P. tricornutum* and *T. pseudonana*.

Phylogenetic analysis of the miRNA precursors

Considering the significance of miRNAs in evolution investigation, the precursor sequences of the predicted diatom miRNAs and the known miRNAs in the same family were aligned and phylogenetically analyzed by ClustalW online to investigate their evolutionary relationships (http://www.clustal.org/).

RESULTS AND DISCUSSION

Identification of the diatom miRNAs

Sequence and structure homologies are the main theory
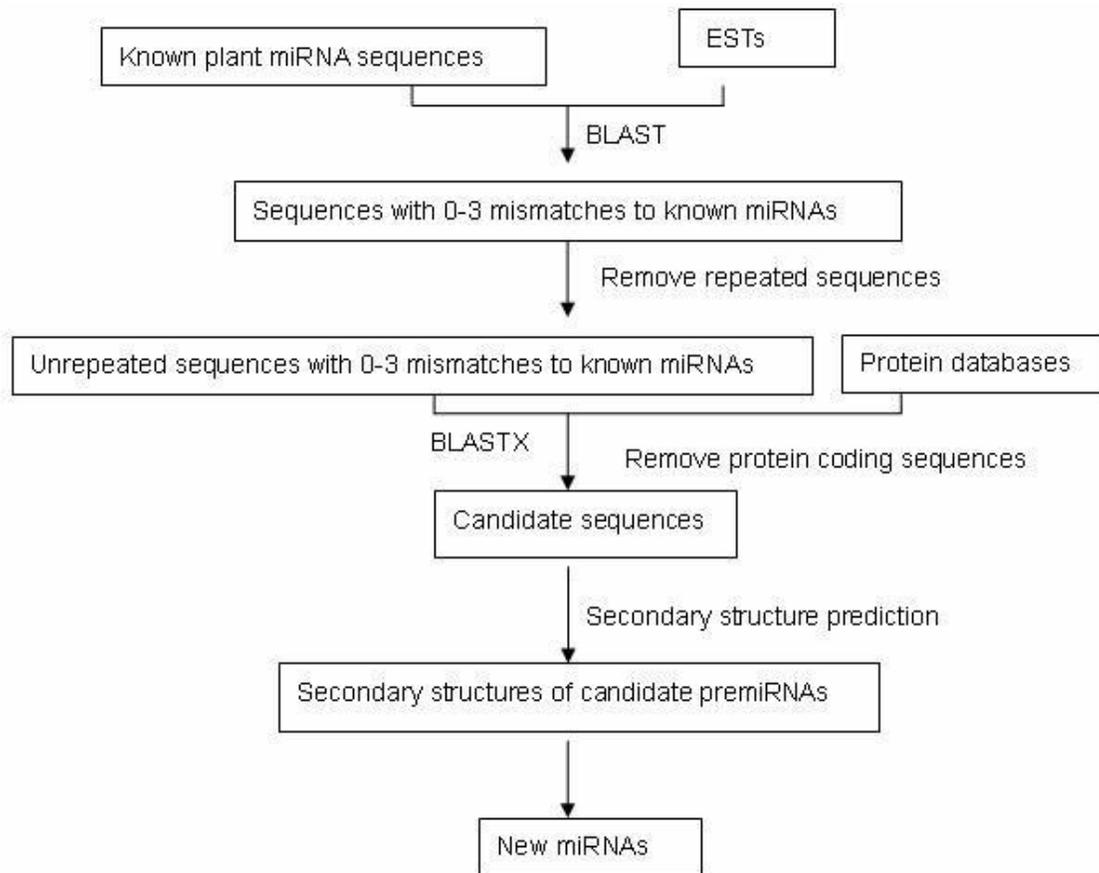
Figure 1. Flowchart of the diatom miRNA prediction.

behind the computer-based approach for miRNAs prediction. As described in materials and methods, after BLASTIN searches, all blasted hits except coding sequences were maintained for secondary structure analysis; only those in line with the screening criteria were selected as candidates. In the end, 5 potential *P. tricornutum* miRNAs belonging to 5 miRNA families and 1 potential *T. pseudonana* miRNA were identified, all being named according to Ambros et al. (2003). Information on predicted diatom miRNAs was listed in Table 1. The length of the 6 predicted miRNAs ranged from 20 to 21 nt, while the predicted precursor sequences ranged in length from 45 to 57 nt, all forming into typical stem-loop structures, with the mature miRNA either on the 5' end or the 3' end (Figure 2). The MFEIs of these hairpin structures were about 0.85, which was thought to be the gold standard to differentiate miRNAs from other ones (Zhang et al., 2006).

Computer-based miRNA identification is becoming more and more important than experimental approaches due to its advantages of low cost and high efficiency (Unver et al., 2009). Considering the taxonomy of algae and the conservation of miRNAs, it's reasonable to use the sequences of known plant miRNAs as the query for BLAST search against the EST databases of the diatoms.

According to Zhang et al. (2006), about 10000 ESTs contained one miRNA, so about 20 miRNAs should be predicted theoretically from the total of 195811 ESTs. Why only 6 were predicted in this work? *In silico* analysis of the diatom genomes for known components of the RNAi pathway has indicated that molecular players involved in RNA silencing in other eukaryotes are only poorly conserved in diatoms (Riso et al., 2009). So it is concluded that there might be some different characteristics for the diatom miRNAs. Understanding these characteristics will be helpful for the improvement of this method and that will rely on the finding of more miRNAs from the diatoms.

Prediction of the diatom miRNA targets

Based on the homology between miRNAs and their target genes in plants, the *P. tricornutum* and *T. pseudonana* EST databases were searched for homology to the selected miRNA sequences with the BLASTIN algorithm for the discovery of miRNA targets. The miRNAs and their candidate targets were listed in Table 2. Only 2 potential targets were identified for the 5 *P. tricornutum* miRNAs and 3 potential targets for the *T. pseudonana* miRNA

Table 1. miRNAs identified in the diatoms by bioinformatics method.

miRNAs	Gene ID	Source	miRNA sequence (5'to 3')	NM/nt	LM/nt	LP/nt	Location	A + U (%)	MFEI
pha-MIR399d	186996709	EST	UGCCAAGAAAGAUUUGCCCCU	3	21	52	3	0.5	0.83
pha-MIR837-5p	144671114	EST	AUCAGUUUCUUGAUCGUUUCA	1	21	53	3	0.64	0.86
pha-MIR1886.2	187018652	EST	UGGGAUGAAAUCUUUGAGUAG	3	21	57	5	0.59	0.85
pha-MIR 2106	144661892	EST	CCGAGGUUUUCUUGAUUCUUU	3	21	56	5	0.48	0.85
pha-MIR 2911	186995880	EST	GGCCGGGGGACUGGGUGGAA	3	20	55	3	0.35	0.84
tha-MIR414	162364940	EST	UCAAUCUCAUCAUCAUCGUCC	2	21	45	5	0.62	1.06

NM: number of mismatch, LM: length of mature miRNAs, LP: length of precursor, MFEI: minimal folding free energy index.
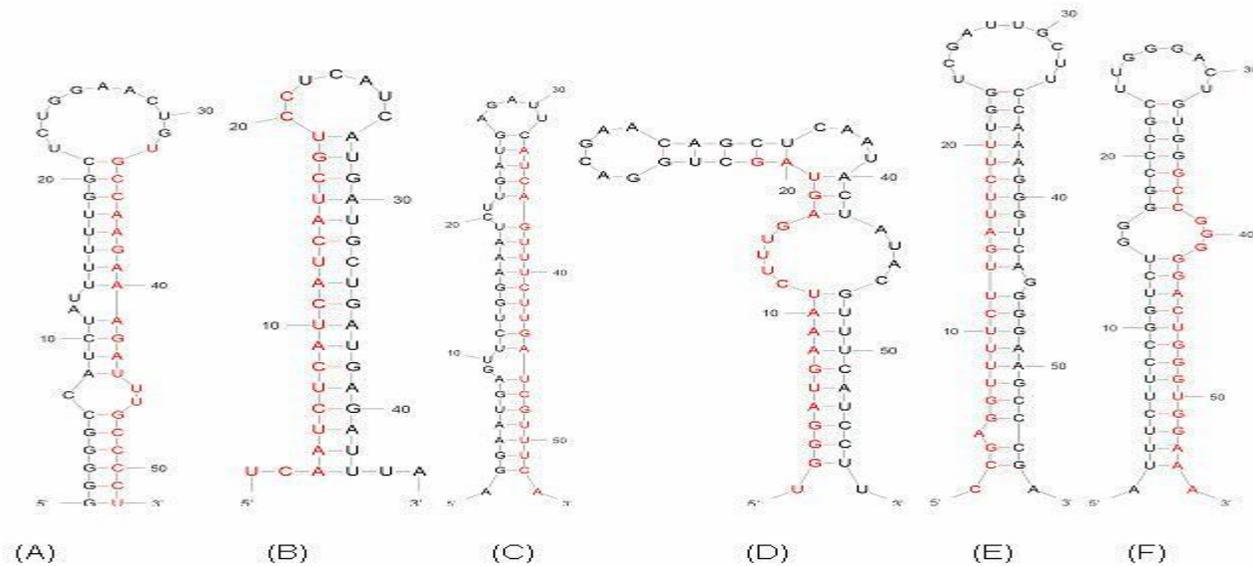
Figure 2. Secondary structures of the diatom pre-miRNAs. Mature miRNA sequences are red. A: pha-MIR399d precursor, B: pha-MIR837-5p precursor, C: pha-MIR1886. 2 precursor, D: pha-MIR 2106 precursor, E: pha-MIR 2911 precursor, F: tha-MIR414 precursor.

Table 2. The potential targets of newly identified miRNAs in the diatoms.

miRNA	Targeted gene	Targeted protein	Possible function
tha-MIR414	162379077	chromosome condensation protein-like protein	cell division
	162368912	nifu-like protein	transcription regulation
	162396859	engrailed-b homeobox protein	transcription regulation
pha-MIR1886.2	187020570	alpha/beta hydrolase fold protein	metabolism
pha-MIR 2911	186995880	hypothetical protein	unknown

Table 3. Comparisons of sequence similarity between the diatom miRNAs and corresponding miRNA family members.

			similarity	
Index	Family	Count	Mature (%)	Pre-miRNA (%)
1	MIR399d	9	< 80	< 36
2	MIR837-5p	1	95	20
3	MIR1886.2	1	85	19
4	MIR 2106	1	85	17
5	MIR 2911	1	85	32
6	MIR414	3	< 90	< 64

Family: the name of miRNA family, count: the number of miRNAs in other species in the same miRNA family, similarity: the similarity between the sequences of the diatom miRNAs and the sequences of other members of the same family.

were identified, too.

Target identification is an effective way to assess and define the putative function for a miRNA in plant and EST data base searching has played a vital role for the discovery of miRNA targets in plants based on the homology between miRNA and its target sequences (Fahlgren et al., 2010). Our prediction of target mRNAs for the 6 miRNA candidates revealed that a broad range of processes including transcription, metabolism and cell division might be regulated by miRNAs (Table 2). The reason why no targets were identified for 2 *P. tricornutum* miRNAs might be related to the volume of the database.

Phylogenetic analysis of the diatom miRNA precursors

Though plant miRNAs are highly conserved among distantly related plant species, both in terms of primary and mature miRNAs (Zhang et al., 2006), variations in pre-miRNA sequences provides the chance to investigate their evolutionary relationships. Sequence comparisons of the members in the same miRNA families showed that the all the mature miRNAs of the diatoms had a high degree of sequence similarity with the other members, but most of the precursor sequences were on the contrary, with sequence similarities lower than 36% (Table 3). That is, the precursor sequences of the miRNAs from the diatoms are poorly conserved, which is different from

other plant species, but consistent with the report of Riso et al. (2009).

Based on the results of sequence comparisons, only the precursor sequences of the members in MIR399d and MIR414 families were phylogenetically analyzed, since there existed just one member in the family of MIR837, MIR1886, MIR2106 and MIR2911, respectively. It could be seen from the phylogenetic trees that the evolutionary relationship of the diatom was somewhat remotely related to other plant species (Figure 3), which was consistent with the studies from fossil, biological, genomic and molecular aspects (Patricia et al., 2006). The finding of miRNAs has provided another kind of genetic material for the evolution investigation of diatoms.

Conclusions

By the bioinformatics method, 6 miRNAs were identified from the EST databases of *P. tricornutum* and *T. pseudonana*, which provided evidence for the first time that miRNAs were just as effective in diatoms as in other eukaryotes and their potential targets prediction indicated their roles in transcription, metabolism and cell division. Alignmental and phylogenetic analyses showed that the miRNA precursor sequences of the diatoms were poorly conserved and the evolutionary relationship was somewhat remotely related to other plant species. Above all, the findings from this study will contribute not only to
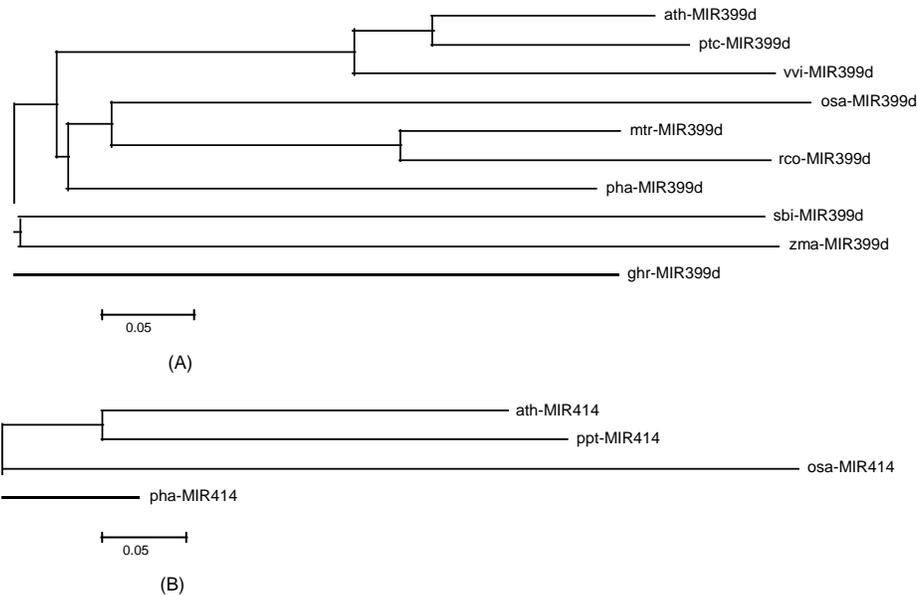
Figure 3. Phylogenetic trees of MIR399d and MIR414 families, A: MIR399d family, B: MIR414 family.

further research of miRNAs features and regulatory mechanisms in diatoms, but also to their evolutionary research by virtue of new molecular tools.

ACKNOWLEDGEMENTS

REFERENCES

Ambros V, Bartel B, Bartel DP (2003). A uniform system for micro RNA annotation. RNA, 9(3): 277-279.

Ambros V, Chen X (2007). The regulation of genes and genomes by small RNAs. Development, 134(9): 1635-1641.

Armbrust EV (2009). The life of diatoms in the world's oceans. Nature. 459(7244): 185-192.

Axtell MJ, Bowman JL (2008). Evolution of plant microRNAs and their targets. Trends Plant Sci., 13(7): 343-349.

Barbato C, Arisi I, Frizzo ME, Brandi R, Da Sacco L, Masotti A (2009). Computational challenges in miRNA target predictions: to be or not to be a true target? J. Biomed. Biotechnol., p. 803069.

Bartel DP (2004). MicroRNAs: genomics, biogenesis, mechanism and function. Cell. 116(2): 281-297.

Baskerville S, Bartel DP (2005). Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes. RNA, 11(3): 241-247.

Bonnet E, Wuyts J, Rouzé P, Van de Peer Y (2004). Detection of 91 potential conserved plant microRNAs in *Arabidopsis thaliana* and *Oryza sativa* identifies important target genes. Proc. Natl. Acad. Sci. USA, 101(31): 11511-11516.

Bozarth A, Maier UG, Zauner S (2009). Diatoms in biotechnology: modern tools and applications. Appl Microbiol Biotechnol., 82(2): 195-201.

Carrington JC, Ambros V (2003). Role of microRNAs in plant and animal development, Sci., 301(5631): 336-338.

Chen X (2004). A microRNA as translational repressor of APETALA2 in *Arabidopsis* flower development, Sci., 303(5666): 2022-2025.

Dagan T, Martin W (2009). Microbiology. Seeing green and red in diatom genomes. Sci., 324(5935): 1651-1652.

De Riso V, Raniello R, Maumus F, Rogato A, Bowler C, Falciatore A (2009). Gene silencing in the marine diatom *Phaeodactylum tricornutum*. Nucleic Acids Res., 37(14):e96.

Ding YF, Zhu C (2009). The role of microRNAs in copper and cadmium homeostasis. Biochem. Biophys. Res. Commun., 386(1): 6-10.

Fahlgren N, Carrington JC (2010). miRNA Target Prediction in Plants, Methods .Mol .Biol., 592: 51-57.

Fornara F, Coupland G (2009). Plant phase transitions make a SPLash. Cell., 138(4): 625-627.

Han Y, Luan F, Zhu H, Shao Y, Chen A, Lu C, Luo Y, Zhu B (2009). Computational identification of microRNAs and their targets in wheat (*Triticum aestivum L.*). Sci. China, Ser. C Life Sci., 52(11): 1091-1100.

He PA, Nie Z, Chen J, Chen J, Lv Z, Sheng Q, Zhou S, Gao X, Kong L, Wu X, Jin Y, Zhang Y (2008). Identification and characteristics of microRNAs from *Bombyx mori*. BMC Genomics., 9: 248.

Kosik KS (2009). MicroRNAs tell an evo-devo story. Nat Rev Neurosci. 10(10): 754-759.

Kroth PG (2007). Genetic transformation: a tool to study protein targeting in diatoms. Methods Mol. Biol., 390: 257-67.

Lu S, Sun Y-H, Shi R, Clark C, Li L, Chiang VL (2005). Novel and mechanical stress-responsive microRNAs in *Populus trichocarpa* that are absent from *Arabidopsis*. Plant Cell, 17(8): 2186-2203.

Maheswari U, Mock T, Armbrust EV, Bowler C (2009). Update of the Diatom EST Database: a new tool for digital transcriptomics. (Database issue): Nucleic Acids Res., 37: D1001-1005.

Mathieu J, Yant LJ, Mürdter F, Küttner F, Schmid M (2009). Repression of flowering by the miR172 target SMZ. PLoS Biol., (7): e1000148.

Nunn BL, Aker JR, Shaffer SA, Tsai S, Strzepek RF, Boyd PW, Freeman TL, Brittnacher M, Malmström L, Goodlett DR (2009). Deciphering diatom biochemical pathways via whole-cell proteomics, Aquat. Microb. Ecol., 55(3): 241-253.

Patricia A, Sims David G, Mann Linda K, Medlin (2006). Evolution of the diatoms: insights from fossil, biological and molecular data, Phycol., 45(4): 361-402.

Pothof J, Verkaik NS, van IJcken W, Wiemer EA, Ta VT, van der Horst GT, Jaspers NG, van Gent DC, Hoeijmakers JH, Persengiev SP (2009). MicroRNA-mediated gene silencing modulates the UV-

induced DNA-damage response. EMBO J., 28(14): 2090-2099.

Reinhart BJ, Weinstein EG, Rhoades MW, Bartel B, Bartel DP (2002). MicroRNAs in plants. Genes. Dev., 16(13): 1616-1626.

Rodriguez RE, Mecchia MA, Debernardi JM, Schommer C, Weigel D, Palatnik JF (2010). Control of cell proliferation in *Arabidopsis thaliana* by microRNA miR396. Development, 137(1): 103-112.

Ruiz-Ferrer V, Voinnet O (2009). Roles of plant small RNAs in biotic stress responses. Annu. Rev. Plant, Biol., 60: 485-510.

Song C, Fang J, Li X, Liu H, Thomas Chao C (2009). Identification and characterization of 27 conserved microRNAs in *citrus*. Planta., 230 (4): 671-685.

Sunkar R, Chinnusamy V, Zhu J, Zhu JK (2007). Small RNAs as big players in plant abiotic stress responses and nutrient deprivation. Trends. Plant. Sci.*,* 12(7): 301-309.

Sunkar R, Jagadeeswaran G (2008). In silico identification of conserved microRNAs in large number of diverse plant species. BMC Plant. Biol., 8: p.37.

Unver T, Namuth-Covert DM, Budak H (2009). Review of Current Methodological Approaches for Characterizing MicroRNAs in Plants. Int. J. Plant. Genomics, pp. 262463.

Vardi A, Thamatrakoln K, Bidle KD, Falkowski PG (2008). Diatom genomes come of age. Genome Biol., 9(12): 245.

Wang JW, Czech B, Weigel D (2009). miR156-regulated SPL transcription factors define an endogenous flowering pathway in *Arabidopsis thaliana*. Cell, 138(4): 738-749.

Xie FL, Huang SQ, Guo K, Xiang AL, Zhu YY, Nie L, Yang ZM (2007). Computational identification of novel microRNAs and targets in *Brassica napus*. FEBS Lett., 581(7): 1464-1474.

Yin Z, Li C, Han X, Shen F (2008). Identification of conserved microRNAs and their target genes in tomato (*Lycopersicon esculentum*). Gene., 414 (1-2): 60-66.

Zhang B, Pan X, Cannon CH, Cobb GP, Anderson TA (2006). Conservation and divergence of plant microRNA genes. Plant J., 46(2): 243-259.

Zhang BH, Pan XP, Cox SB, Cobb GP, Anderson TA(2006). Evidence that miRNAs are different from other RNAs. Cell. Mol. Life. Sci., 63(2): 246-254.

Zhang BH, Pan XP, Wang QL, Cobb GP, Anderson TA (2005). Identification and characterization of new plant microRNAs using EST analysis. Cell Res., 15(5): 336-360.

Zhang W, Luo Y, Gong X, Zeng W, Li S (2009). Computational identification of 48 potato microRNAs and their targets. Comput, Biol. Chem., 33(1): 84-93.

Zuker M (2003). Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res., 31(13): 3406-3415.