In silico Analysis of MicroRNAs in *Spinacia oleracea* Genome and Transcriptome

Bihter Avsar*, Danial Esmaeili Aliabadi
Sabanci University, Faculty of Engineering and Natural Sciences, Istanbul, Turkey.

* Corresponding author. Email: bihteravsar@sabanciuniv.edu

Manuscript submitted October 10, 2016; accepted January 13, 2017.
doi: 10.17706/ijbbb.2017.7.2.84-92

Abstract: Plant microRNAs (miRNAs) are small non-coding RNAs, about 21-24 nucleotides, which have important regulatory roles in growth, development, metabolic and defense processes. These critical elements regulate pathways either by inducing translational repression or messenger RNA (mRNA) decay. With the advent of the next-generation sequencing technologies and newly developed bioinformatics tools, the identification of microRNA studies by computational methods have been increased. Thus, the sequencing information provides us information for mining some known and unknown miRNAs in plants. In this study, we predict 34 putative miRNAs from *Spinacia oleracea* genome and two putative miRNA families from spinach transcriptome by using homology-based conservation method. RepeatMasker program is utilized to mask and eliminate five miRNA families out of 34 putative miRNA families from spinach genome. Finally, we analyze the targets of putatively identified miRNAs and their representation of genes (the copy number of each miRNA) throughout the genome.

Key words: *In silico* prediction, microRNA, miRNA, spinach, *Spinacia oleracea.*

1. Introduction

Increasing world population, the drastic climate change, the threat of biotic stress and abiotic stress factors on the plants and the scarcity of arable lands have brought major concerns about the food security and its sustainability in the future; therefore, new agricultural technologies should be developed to prevent the worst-case scenario in the future [1]. With the advent of the next-generation sequencing (NGS) technologies, complex genomes of organisms can be unraveled [2]. Although new sequencing technologies are immature yet and not optimal compare to previous methods, such as Sanger sequencing, they reduce the process time, cost and the effort required [2]. Thanks to the impact of NGS technologies, most of the biotechnological methods are developed and the crops are improved [3].

MicroRNAs (miRNAs) are small, about 21-24 nucleotides, endogenous non-coding RNAs that play various roles in plants and animals. They regulate the miRNA-gene expression at the post-transcriptional level, and they are primarily located in intergenic regions of plant genomes [3]. MicroRNAs are derived from the stem-loop structure, and are modified by some specific enzymes. Plant microRNAs control the expression of genes encoding various transcription factors, stress-responsive elements, and the other proteins which have roles in growth, development and physiological properties [4]. Rapid growth in miRNA identification studies and related tools that use genomic data exposes the interest in this line of research among researchers from academia as well as practitioners from industries. Identifying miRNAs via computational methods is qualified by reaching to the successful means, and some new miRNA families were validated experimentally. These experimentally identified miRNAs have roles on abiotic stresses due to drought, salinity, heat, cold, phosphorous deficiency or biotic stresses [4]. Currently, computational miRNA prediction is based on two methods: 1) Homology-based for conserved miRNA identification 2) other algorithms which use Support Vector Machine by setting some characteristics for pre-miRNA structure [5]. In our study, we used the 'homology-conserved' method to predict some putative miRNAs via employing in-house Perl scripts [6].

Spinach (*Spinacia oleracea*) is a valuable plant to study in different research areas including physiology, molecular biology and biochemistry [7]. For instance, Sklensky and Davies [8] found a relationship between plant senescence and resource partitioning to male and female flowers of *Spinacia oleracea*. In addition to this, cytogenetic studies have been performed on spinach to understand the sex determination mechanism [9]. In 2014, Dohm *et al*. [10] sequenced spinach genome with the *Beta vulgaris* species and they identified a significant number of genes affecting agronomically important traits. Yan *et al*. [11] showed the transcriptome and gene expression profile of spinach under the heat stress to elucidate its cold tolerance; they identified candidate genes so gene-gene interaction pathways might be found to understand the heat resistance mechanism in spinach in the future.

In this article, we predict some putative miRNAs in spinach genome and transcriptome. We also investigate the targets of these predicted miRNAs. We show the copy number of those miRNA genes as representatives of each miRNA families *in silico*. RepeatMasker program is used to observe repetitive miRNAs throughout the genome.

2. Materials and Methods

2.1. Reference miRNAs and Spinach Dataset

Currently, mature miRNA sequences (8,496 sequences and 73 plant species) are available in miRBase release 21 [12]. miRBase corresponds to 4,802 unique mature miRNA sequences, and they were used as queries in homology-based *in silico* miRNA identification. Also, spinach (Viroflay genotype) genome data has been retrieved from a publicly available website[1] [10].

2.2. Transcriptome Dataset

Spinacia oleracea (Viroflay genotype) raw RNA reads (SP78, SP82 and SP90)[2] have been downloaded, and then assembled by using Trinity Genome Guided Transcriptome Assembly software based on the manual's instructions[3] [13]. Finally, three different assembled transcriptomes for SP72, SP80 and SP90 have been created. We used these assemblies separately to predict putative miRNA families and the predicted putative miRNAs were called as "transcriptomic miRNAs".

2.3. *In silico* miRNA Identification Based on Homology Conserved Method

For the prediction, we employed two previously developed, in-house Perl scripts: SUmirFind and SUmirFold, as described in details in [6]. In the first step of homology-based miRNA prediction, BLAST+ stand-alone toolkit, version 2.2.25 [14] was used for the detection of database sequences with homology (mismatch cutoff parameter set to ≤ 3) that previously were known plant mature miRNAs [5], [15]. In order to obtain secondary structures of predicted miRNAs, UNAFold version 3.8 was used with optimized parameters to include all possible stem-loops generated for each miRNA query. Hairpins with multi-branched loops, inappropriate DICER cut sites at the ends of the miRNA-miRNA* duplex, or mature

[1] https://www.ncbi.nlm.nih.gov/assembly/GCA_000510995.2/

[2] ftp://www.spinachbase.org/pub/spinach/transcriptome/fastq

[3] https://github.com/trinityrnaseq/trinityrnaseq/wiki

miRNA sequence portions at the head of the pre-miRNA stem-loop were eliminated manually.

2.4. Repeat Masker Analysis of Putative miRNAs in Spinach Genome

Masking and identification of repetitive elements were performed by a semi-automated pipeline, RepeatMasker version 4.0.6 (www.repeatmasker.org) at default settings with Cross-Match[4] as an alignment algorithm. MIPS-REdat_ALL v9.3[5] was used as the repeat library [16].

2.5. Representative (Copy Number) miRNAs in Spinach Genome

To avoid over-representation, we eliminated the repeated identical miRNAs which might be caused by similar query miRNA stem loop sequences.

2.6. Expressed Sequence Tag (EST) Analysis and Target Annotation of Predicted Genomic miRNAs

For EST analysis, the pre-miRNA sequences were retrieved, and duplicate sequences were removed to prevent over-representation. By using BLAST+ stand-alone toolkit, version 2.2.25 [14] pre-miRNA sequences were blasted to Spinach unigene sequences[6]. We downloaded 72,148 unigene sequences and the strict criteria were used for the analysis of the only miRNA families who had hits above the threshold as 98% identity and 99% query coverage.

For the target annotation analysis, mature sequences were identified, and duplicates were removed. By using online web tool, psRNA[7], we obtained a query file which had targets. These targets were blasted to Spinach unigenes. Then, the results file has been downloaded and used as a query for gene ontology analysis. Blast2Go[8] online software was used [17] for gene ontology analysis. Additionally, we searched predicted mature miRNA sequences in miRBase [12] database to confirm their experimentally validated targets.

3. Results and Discussions

3.1. Putative miRNAs through *Spinacia oleracea* Genome and Its Transcriptome

Lower Minimal Folding-Free Energy (MFE) values show high stability of predicted miRNAs. We calculate Minimal Folding Free-Energy Index (MFEI) by using MFE and GC% for genomic spinach miRNAs and the corresponding statistics are reported in Table 1. Minimal Folding Free-Energy Index (MFEI) differentiates miRNAs with typically higher MFEIs (> 0.67) from other types of cellular ssRNAs for which MFEIs were previously characterized: transfer RNAs (0.64), ribosomal RNAs (0.59), and mRNAs (0.62–0.66) (Fig.1.)(Supplementary Data 1) [18]. We could predict 34 putative miRNAs in spinach genome. miR5175, miR845, miR8766, miR5181 and miR1130 families were masked by RepeatMasker. RepeatMasker scans repetitive elements in query sequences based on the chosen RepeatMasker library. All putative predicted and masked miRNAs are shown in Table 2. The predicted miRNA structure and sequence is represented in Fig. 1.

We could not detect any predicted miRNA families in SP78 transcriptome assembly; however in SP82 and SP90 transcriptome assemblies, listed miRNA families are identified:

- SP82: miR167, miR5769, and miR6270
- SP90: miR6270.

[4] www.phrap.org/phredphrapconsed.html
[5] http://pgsb.helmholtz-muenchen.de/plant/recat/
[6] http://www.spinachbase.org/cgi-bin/spinach/download.cgi
[7] http://plantgrn.noble.org/psRNATarget/target
[8] https://www.blast2go.com/

Table 1. Statistics Values for Spinach Genomic miRNAs

	Minimum	Average	Median	Maximum
Sequence length of pre-miRNAs	98	150±50.5nt	134	337
Sequence length of mature miRNAs	20	21.2±0.93	21	24
Minimal folding-free energy (MFE)	-93.3	-58.61±13.86	-56.3 kcal/mol	-25.7
GC%	26.11	38.88±6.60	40	52.29
Minimal Folding Free-Energy Index (MFEI)	0.68	1.05±0.17	1.01	1.63

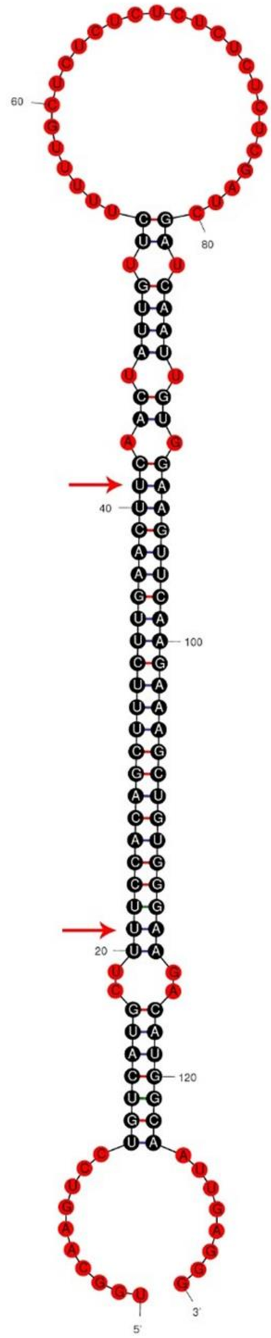

Fig. 1. Identified pre-miRNA stem loop structure of selected miRNA on spinach genome. Mature miRNA start and end points are showed designated by arrows. Structures are predicted using UNAFold program-an implementation of Zuker algorithm.

To verify our results, MFE, GC%, and MFEI values of transcriptomic miRNAs have been checked, and they supported the aforementioned criteria (Supplementary Data 2, Supplementary Data 3).

Table 2. Predicted Putative miRNAs from Unmasked and Masked Spinach Genome

Unmasked genome miRNAs		Masked genome miRNAs	
miR160	miR393	miR160	miR393
miR162	miR399	miR162	miR399
miR172	miR403	miR172	miR403
miR319	miR827	miR319	miR827
miR396	miR395	miR396	miR395
miR156	miR5168	miR156	miR5168
miR159	miR5174	miR159	miR5174
miR167	miR5175	miR167	miR1863
miR394	miR845	miR394	miR1436
miR408	miR1863	miR408	miR5049
miR157	miR8766	miR157	miR1511
miR164	miR1436	miR164	miR535
miR165	miR5049	miR165	
miR166	miR1511	miR166	
miR169	miR535	miR169	
miR170	miR5181	miR170	
miR171	miR1130	miR171	

3.2. Representation of miRNAs in *Spinacia oleracea* Genome

For this section, we use unmasked genomic data to find representatives of each miRNA families in spinach genome. According to the analysis, miR169 families are highly represented whereas miR1130, miR151, miR170, miR403, miR535, miR845 and miR876 families have lower representation (Fig.2.). Low representations of miRNA families are also included in the analysis because they might be 'young-miRNAs'. On the other hand, the highest number of hits might be repetitive elements since most of the transposable elements have been domesticated into microRNA genes and they have high number of copies throughout the genome [19].

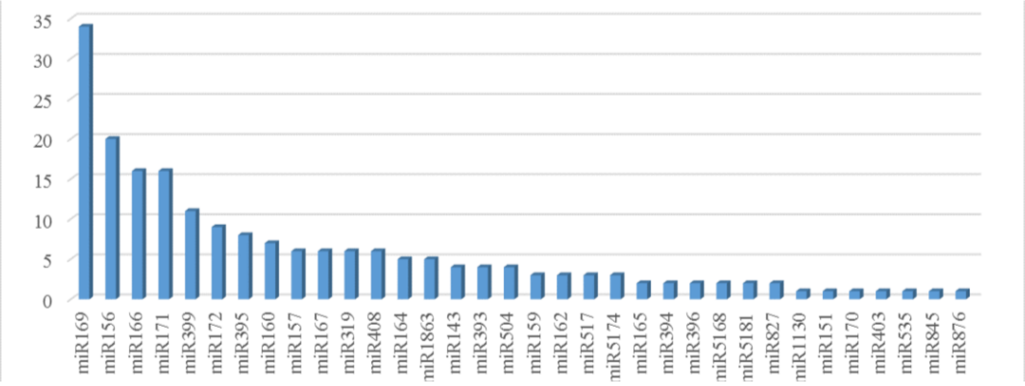

Fig. 2. Representation of predicted putative miRNAs on spinach genome.

3.3. Target Prediction and Gene Ontology Analysis

All predicted putative miRNAs were searched through the miRBase database to understand whether they have experimentally validated targets [8]. According to the results, miR156, miR157, miR159, miR160, miR162, miR164, miR165, miR166, miR167, miR169, miR170, miR171, miR172, miR319, miR393, miR394, miR395, miR396, and miR399 have experimentally validated targets (Table 3). Most of these targets are transcription factors, promoter-binding proteins, and F-box proteins. As depicted in Fig. 3, top species that have highly similar genes to spinach organism are specified by Blast2Go.

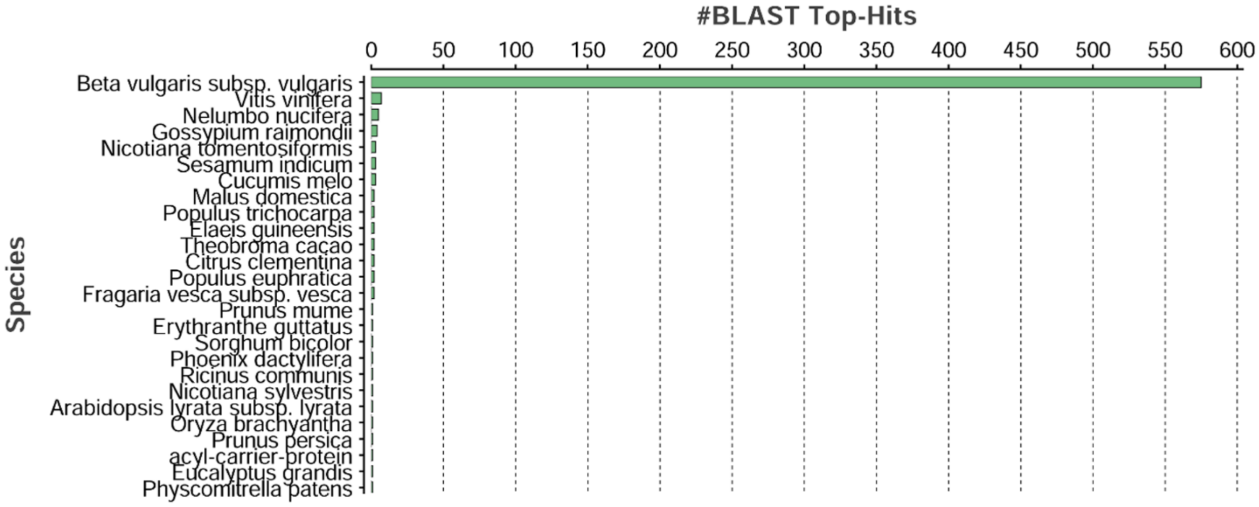

Fig. 3. Top-Hit species are shown after blast results to spinach genome.

Based on these results, *Beta vulgaris* have the most similar genes to spinach genes. In Supplementary Table 1, potential silencing mechanism of putative spinach miRNAs and the predicted functions of their target genes are shown. Blast2Go is utilized to visualize molecular functions, biological processes, and cellular components of identified targets from predicted miRNAs (Fig. 4.).

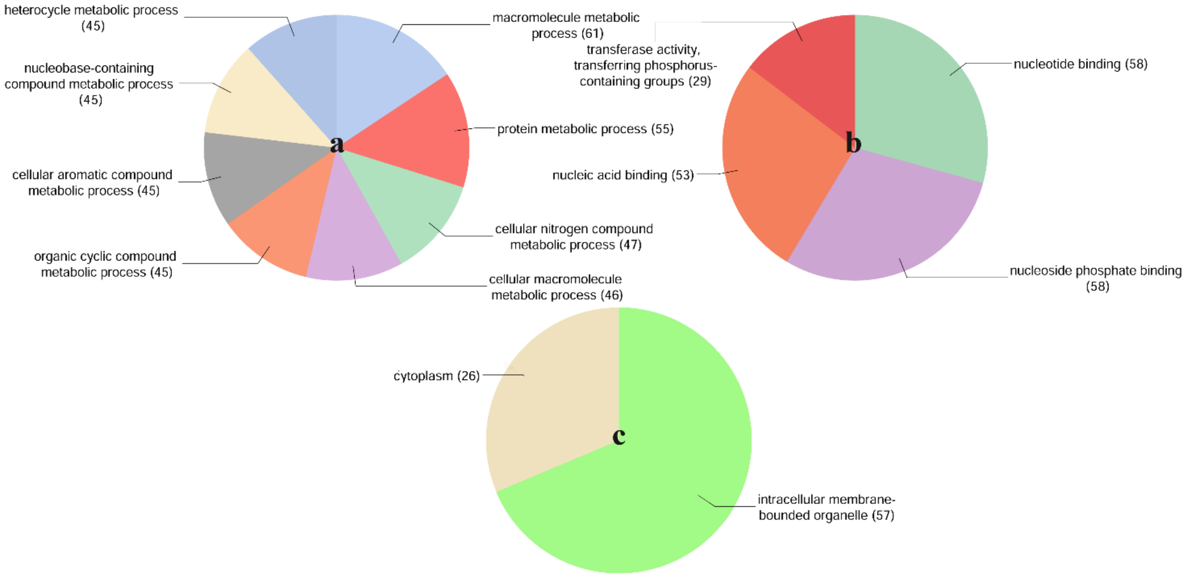

Fig. 4. Target annotation charts of putative spinach miRNAs based on GO analysis are depicted as a.) biological processes, b.) molecular function and c.) cellular component.

EST analysis results show that genomic miR1863 and miR827 families are expressed in spinach genome.

The remaining predicted miRNAs may also be transcribed, but they may not been found in the current unigene file that we use as a database.

Table 3. Experimentally Validated Target Proteins of Predicted miRNAs in miRBase

miRNA	Experimental Targets
miR156	Squamosa-promoter Binding Protein (SBP) box.
miR157	Squamosa-promoter Binding Protein (SBP) box.
miR159	MYB and TCP transcription factors.
miR160	Auxin response factor proteins.
miR162	DICER-LIKE 1 (DL1) proteins.
miR164	NAC domain containing proteins such as Cup-Shaped Cotyledon 2 (CUC2).
miR165	HD-Zip transcription factors including Phabulosa (PHB) and Phavoluta (PHV).
miR166	HD-Zip transcription factors including Phabulosa (PHB) and Phavoluta (PHV).
miR167	Auxin response factors.
miR169	CCAAT Binding Factor (CBF) and HAP2-like transcription factors.
miR170	GRAS domain or SCARECROW-like proteins.
miR171	GRAS domain or SCARECROW-like proteins.
miR172	APETALA2-like transcription factors.
miR319	TCP genes for cleavage.
miR393	F-box proteins and bHLH transcription factors.
miR394	F-box proteins.
miR395	ATP sulphurylases.
miR396	Growth Regulating Factor (GRF) transcription factors, rhodenase-like proteins, and kinesin-like protein B.
miR399	Phosphatase transporter.

4. Conclusions

The advent of modern sequencing technologies and computational methods help researchers to filter redundant data for better understanding of plant genomes. To the best of authors' knowledge, the presented paper is the first study that employed computational methods to identify miRNA elements on the spinach genome. Since spinach has important agronomical properties and sex determination mechanism, identification of its miRNA repertoire may provide some clues about the pathways. Our findings may also help researchers to understand the regulatory roles of putative miRNAs in other spinach accessions which show genetic diversities between each other and those which was analyzed by some molecular markers [20].

For the future studies, widely distributed and highly conserved miRNA families including miR169, miR156, miR166 and miR171 families should be experimentally validated. These miRNAs are known as important elements in different mechanisms ranging from abiotic stress tolerance to seed development; specifically, miR169 families, which are highly represented in spinach genome based on our findings, were shown to be up-regulated under drought and cold stresses in Arabidopsis [21]-[24]. Furthermore, performing evolutionary studies for spinach's relatives to understand their similarities/differences based on the miRNA repertoires and the functions of these putative miRNAs inside the organisms are valuable.

References

[1] Liu, Q., & Chen, Y. Q. (2010). A new mechanism in plant engineering: the potential roles of microRNAs in molecular breeding for crop improvement. *Biotechnol. Adv., 28*, 301-307.

[2] Bolger, M. E., *et al*. (2014). Plant genome sequencing — Applications for crop improvement. *Curr. Opin. Biotechnol., 26*, 31-37.

[3] Tang, G. (2010). Plant microRNAs: An insight into their gene structures and evolution. *Semin. Cell Dev Biol., 21*,782-789.

[4] Rogers, K., & Chen, X. (2013). Biogenesis, turnover, and mode of action of plant microRNAs. *The Plant Cell., 25,* 2383-2399.

[5] Zhang, B., Pan, X., & Wang, Q. (2006). Computational identification of microRNAs and their targets. *Comp. Bio. Chem., 30*, 395-407.

[6] Lucas, S. J., & Budak, H. (2012). Sorting the wheat from the chaff: identifying miRNAs in genomic survey sequences of *Triticum aestivum* chromosome 1AL. *PLOS ONE., 7*, 1-11.

[7] Schmitz-Linneweber, C., *et al.* (2001). The plastid chromosome of spinach (*Spinacia oleracea*): Complete nucleotide sequence and gene organization. *Plant Mol. Biol., 45,* 307-315.

[8] Sklensky, D. E., & Davies, P. J. (2011). Resource partitioning to male and female flowers of *Spinacia oleracea* L. in relation to whole-plant monocarpic senescence. *J. Exp. Bot., 62,* 4323-4336.

[9] Iizuka, M., & Janick, J. (1962). Cytogenetic analysis of sex determination in *Spinacia oleracea*. *Genetics, 47*, 1225-1241.

[10] Dohm, J. C., *et al*. (2014). The genome of the recently domesticated crop plant sugar beet (*Beta vulgaris*). *Nature, 505,* 546-549.

[11] Yan, J., *et al*. (2016). De novo transcriptome sequencing and gene expression profiling of spinach (*Spinacia oleracea* L.) leaves under heat stress. *Sci. Rep., 6*, 1-10.

[12] Kozomara, A., & Griffiths-Jones, S. (2013). miRBase: Annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res., 42,* D68-D73.

[13] Grabherr, M. G., *et al*. (2011). Full-length transcriptome assembly from RNA-seq data without a reference genome. *Nat. Biotechnol., 29*, 644-652.

[14] Camacho, C., *et al*. (2009). BLAST+: architecture and applications. *BMC Bioinform., 10*, 1-9.

[15] Avsar, B., & Aliabadi, D. E. (2015). Putative microRNA analysis of the kiwifruit *Actinidia chinensis* through genomic data. *Int. J. Life Sci. Biotechnol. Pharma Res., 4*, 96–99.

[16] Nussbaumer, T., *et al*. (2013). MIPS PlantsDB: A database framework for comparative plant genome research. *Nucleic Acids Res., 41*, D1144-D1151.

[17] Conesa, A., & Götz, S. (2008). Blast2GO: A comprehensive suite for functional analysis in plant genomics*. Int. J. Plant Genomics., 2008*, 1-12.

[18] Schwab, R., *et al*. (2005). Specific effects of microRNAs on the plant transcriptome. *Dev. Cell., 8,* 517-527.

[19] Li, Y., *et al*. (2011). Domestication of transposable elements into microRNA genes in plants. *PLOS ONE., 6,* 1-13.

[20] Avşar, B. (2011). Genetic diversity of Turkish spinach cultivars. Master disssertation, Izmir Institute of Techonology, Izmir.

[21] Li, W. X., *et al*. (2008). The Arabidopsis NFYA5 transcription factor is regulated transcriptionally and posttranscriptionally to promote drought resistance. *The Plant Cell*, *20*, 2238-2251.

[22] Zhou, X., *et al.* (2008).Identification of cold-inducible microRNAs in plants by transcriptome analysis. *Biochim. Biophys. Acta (BBA)-Gene Regulatory Mechanisms*, *1779*, 780-788.

[23] Zhao, B., *et al.* (2009). Members of miR-169 family are induced by high salinity and transiently inhibit the NF-YA transcription factor. *BMC Mol. Biol.*, *10*, 29-39.

[24] Lee, H., *et al.* (2010). Genetic framework for flowering-time regulation by ambient temperature-responsive miRNAs in Arabidopsis. *Nucleic Acids Res.*, *38*, 3081-3093.

Bihter Avsar received her B.Sc. degree in genetics and bioengineering in Yeditepe University, Istanbul and she completed her M.Sc. in molecular biology and genetics area in Izmir Institute of Technology. She is currently a Ph.D. candidate in Biological Sciences and Bioengineering Department in Sabanci University. Her research interests are bioinformatics, computational biology, genomics, proteomics, biotechnology, plant genetics and molecular biology.

Danial Esmaeili Aliabadi is a postdoctoral research fellow at Sabanci University. He has been participated in RoboCup international competitions from 2009 to 2011 and received many national and international awards. His research interests include artificial intelligence, agent-based simulation, optimization, game theory, data mining, visualization, and machine learning. He holds Ph.D. in Industrial Engineering from Sabanci University, Turkey.