**Bioinformation**
www.bioinformation.net

open access
**Hypothesis**

# High GC content: Critical parameter for predicting stress regulated miRNAs in *Arabidopsis thaliana*

**Akaash Kumar Mishra, Seep Agarwal, Chakresh Kumar Jain, Vibha Rani***

Department of Biotechnology, Jaypee Institute of Information Technology University, NOIDA 201307, India; Viba Rani – Email: Vibha.rani@jiit.ac.in; *Corresponding author

Received June 09, 2009; Revised July 25, 2009; Accepted September 11, 2009; Published October 11, 2009

**Abstract:**
Plants like *Arabidopsis thaliana* are convenient model systems to study fundamental questions related to regulation of the stress transcriptome in response to stress challenges. Microarray results of the *Arabidopsis* transcriptome indicate that several genes could be upregulated during multiple stresses. High-salinity, drought, and low temperature are three common environmental stress factors that seriously influence plant growth and development worldwide. Recently, microRNAs (miRNAs) have emerged as a class of gene expression regulators that have also been linked to stress responses. However, the relationship between miRNA expression and stress responses is just beginning to be explored. Here we have computationally analyzed 123 non redundant miRNA sequences reported for *Arabidopsis thaliana*, including 17 miRNA sequences which were reported to be stress regulated in literature. A significant increase in the GC content of stress regulated miRNA sequences was observed which further extends the view that miRNAs act as ubiquitous regulators under stress conditions. GC content may also be considered as a critical parameter for predicting stress regulated miRNAs in plants like *Arabidopsis thaliana.*

**Keywords:** *Arabidopsis thaliana*, Stress, miRNAs, transcriptome, miRBase, GC content, Bioinformatics

## Background:

Our understanding of the regulation of functional genes responsive to stress signals is still nascent. Multiple signalling pathways regulate the stress responses of plants and there is significant overlap between the patterns of gene expression that are induced in plants in response to different stresses. Many genes induced by stress challenges, including those encoding transcription factors, have been identified and some of them have been shown to be essential for stress tolerance **[1]**. Many studies have also revealed some of the complexity and overlap in the responses to different stresses and are likely to lead to new ways to enhance crop tolerance to disease and environmental stress. MicroRNAs (miRNAs) are a highly conserved class of small noncoding RNAs that regulate gene expression by posttranscriptional degradation or translational repression **[2]**. They play important roles in multiple biological and metabolic processes, including developmental timing, signal transduction, cell maintenance, differentiation and diseases **[3, 4]**. Recently, there has been strong evidence leading to the proposal that miRNAs are hypersensitive to abiotic or biotic stress as well as to diverse physiological processes. More and more evidence has shown that gene silencing is widely adopted in plant immunity. In the past, studies often focused on transposon or siRNA-mediated RNA silence **[5, 6]**. Since miRNAs and siRNAs share many features in common, it is supposed that miRNAs may also be involved in silencing invaders. This was supported by the observation that siRNAs functioned as miRNAs and miRNAs interacted with mRNA in the same way as siRNAs **[7]**. A family of Arabidopsis mRNAs encoding SCARECROW-LIKE (SCL) transcription factors is cleaved by an RNAi-like process directed by miR171 **[8]**. In plant embryo extracts, an endogenous miRNA that lacks perfect complementarity to its RNA targets acts as a siRNA **[9]**. In other words, the data reveals an interchangeable functional role between miRNA and siRNA. Plant virus-derived small RNAs in the gene silencing (VIGS) process were generally considered to be siRNAs. The prevalence of imperfect hairpin structure prompts a re-evaluation of their biochemical nature. In fact, many of these molecules might be akin to miRNAs, because their hairpins have greater similarity to miRNA precursors than to the perfect dsRNAs that produce siRNAs **[10, 11]**. Plant virus infections resulted in a dramatic increase in miRNA whereas virus infected vertebrate cells increased siRNA content **[12]**.

There have been many studies to identify plant miRNAs and numerous miRNAs have been discovered in *Arabidopsis thaliana*. Currently, a variety of biochemical, molecular, and bioinformatics approaches and technologies have been developed for miRNA analysis and detection. Using tiling path microarray analysis as a tool, it is now possible to perform high-throughput profiling of the expression of all the known miRNAs to examine their expression profiles under different environmental stresses **[13]**. In a recent report the effects of 117 miRNAs under high-salinity, drought, and low-temperature stress conditions were analyzed using miRNA chips representing nearly all known miRNAs cloned or identified in *Arabidopsis thaliana* **[14]**. Seventeen stress-inducible miRNAs were detected and the results were further confirmed by detecting their expression patterns and analyzing the *cis*-regulatory elements in their promoter sequences **[14]**. In our work, we analyzed the nucleotide base frequencies at each position of the 123 miRNA sequences that have been reported for Arabidopsis. These 123 sequences obtained from the miRBase include the 117 sequences that were analyzed by Han-Hua L. and coworkers **[14]** and 6 newly reported miRNA sequences in *Arabidopsis thaliana*. We also calculated the base density for each nucleotide in these 123 sequences. This same approach was then used for analyzing groups of miRNA sequences which have been reported to regulate 3 types of abiotic stresses – high salinity, drought and low temperature. A graphical comparison and trend analysis was done to observe differences in frequencies of nucleotide bases between the stress-regulated miRNAs and all miRNA sequences as a whole.

## Methodology:

The microRNA sequence database miRBase was used to create a dataset of 123 non redundant miRNA sequences reported in *Arabidopsis thaliana* (see supplementary material). We grouped the miRNA sequences into 5 groups, 1st containing all the 123 sequences, 2nd containing all the sequences induced by high salinity stress, 3rd by drought stress, 4th by cold stress and 5th by all the three stresses. Further we calculated the relative frequency of each nucleotide base at each position in the miRNA sequence in each of the groups.

### miRBase: the microRNA sequence database

The miRBase Sequence database is freely accessible primary repository for published microRNA (miRNA) sequence and annotation data. miRBase provides a user-friendly web interface for miRNA data, allowing the user to search using key words or sequences, trace links to the primary literature referencing the miRNA discoveries, analyze genomic coordinates and context, and mine relationships between miRNA sequences. miRBase also provides a confidential gene-naming service, assigning official miRNA names to novel genes before their publication **[15]**.

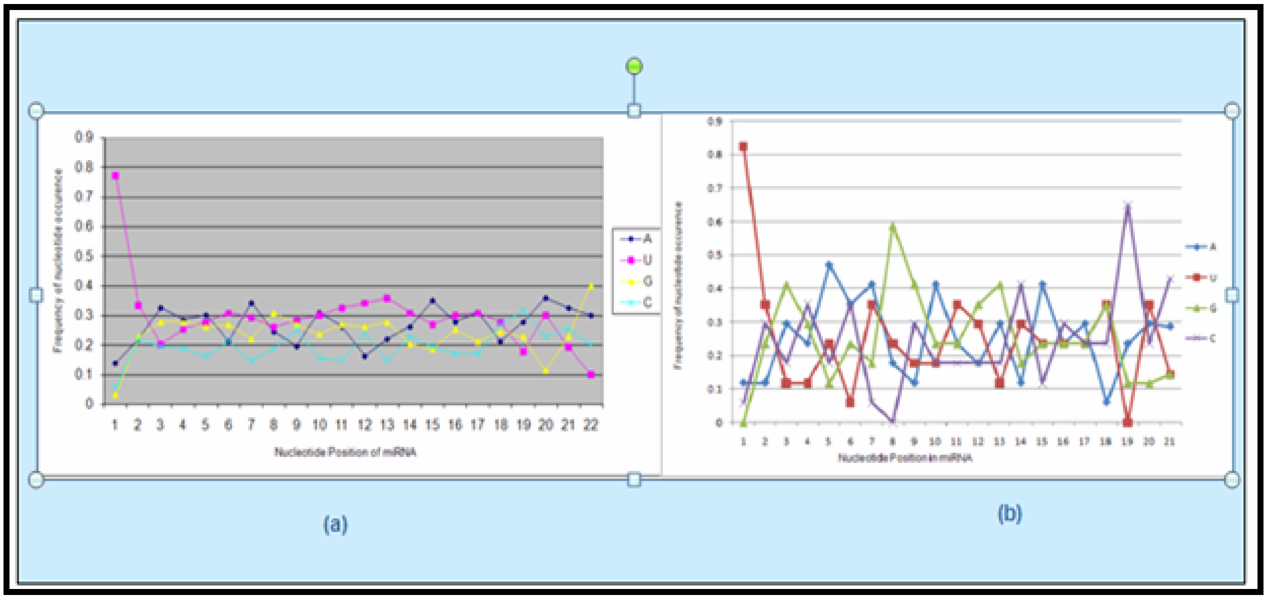

**Figure 1:** Nucleotide frequency distribution of miRNAs: (a) The nucleotide position wise AUGC distribution of all miRNAs. (b) The nucleotide position wise AUGC distribution for stress regulated miRNAs. The degree of randomness in the position wise nucleotide frequency distribution is more in the stress regulated miRNAs.

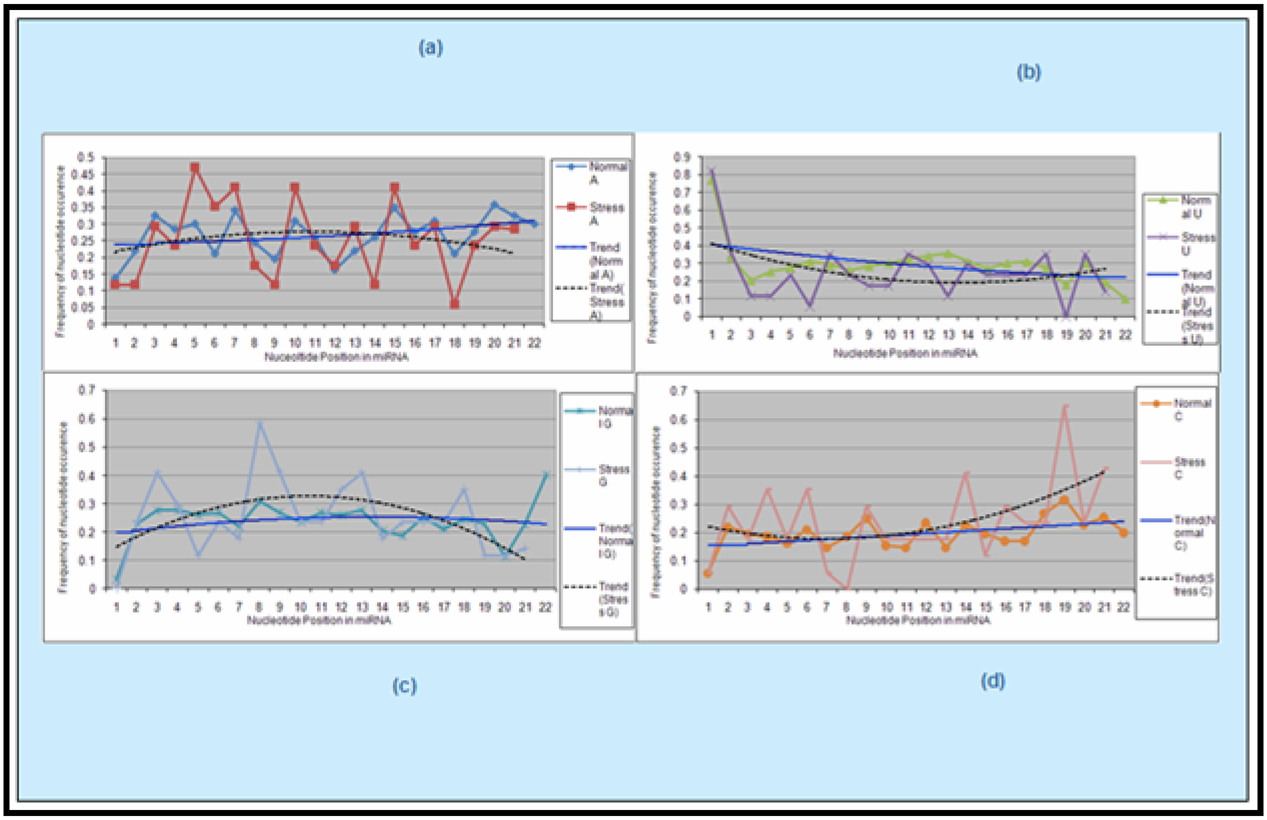

**Figure 2:** Position based trend analysis of nucleotide frequency (a), (b), (c), (d) show the comparison between stress regulated and normal miRNAs for frequency distribution of nucleotide A, U, G and C respectively at each position of the corresponding sequences. The group of normal sequences corresponds to the 1st group of sequences which includes all 123 sequences of the dataset consisting of both the stress regulated sequences as well those not regulated by stress.

**Calculation of position specific nucleotide frequencies:**
All sequences were divided into no. of cells corresponding to their length with each nucleotide in each cell of an excel sheet. The sequences were then aligned in the above mentioned five groups respectively. The following function was used to count the no. of each nucleotide at each position.

**COUNTIF (range, criteria):** Counts the number of cells within a range that meet the given criteria.

**Calculation of position specific relative frequencies:**
The relative frequency of each nucleotide at each position of the sequence was calculated by dividing the absolute frequencies (position specific nucleotide frequencies) by the no. of sequences having a nucleotide base at that specific position. To find out the no. of sequences having a nucleotide base at each position we sum up the absolute frequencies of all the nucleotide bases at each position by SUM (number1, number2...) - Adds all the numbers in a range of cells.

**Finding out nucleotide content in a group of sequences:**
The absolute frequency of each nucleotide at all positions of a group of sequences was summed up using the above mentioned SUM function. The total no. of sequences at each position was also summed up using the same function. The percentage content of each nucleotide in a given group of sequences was calculated by the following formula.

**Percentage content of N = 100 * Sum of absolute frequencies of N / Sum of total no. of sequences at each position.**
Line graphs were plotted to compare the variations in the pattern of position wise nucleotide frequency distribution between stress regulated miRNAs and all miRNAs (**Figure 1**). Then TREND LINEs were added to these line graphs to compare the trend of nucleotide frequency distribution for each nucleotide base in stress regulated sequences and all miRNAs (**Figure 2**).

**Discussion:**
There is a significant increase in the GC content of the sequences regulated by the three stresses. The average GC content of Arabidopsis thaliana genome has been reported as just 36% due to the large intronic regions present which contain large stretches of A-T base pairs. The average GC content of all miRNAs reported for *Arabidopsis thaliana* is calculated to be 43% by our analysis. On the other hand the GC content of stress induced sequences calculated by our method is significantly higher 51-52% (**Table 1 in supplementary material**).

Within a long region of genomic sequence, genes are often characterized by having a higher GC-content in contrast to the background GC-content for the entire genome. The high GC-content content is also a feature of genes coding for transcriptional activators or repressors which regulate gene expression under various stress condition. The promoter sequences of many stress regulated genes have also been reported to have a high GC content **[16]**. Since the miRNAs which were induced by stress were also found to have a higher GC content like the coding regions, it may be hypothesized that the GC content may play a significant role in the interaction of these miRNAs with their target genes for regulating gene expression under stress conditions. The high GC content may be significant because of two reasons – firstly the higher GC ratio gives a higher chance of complementarities with the coding regions also having a higher GC ratio, secondly since GC pair is bound by 3 hydrogen bonds, these bonds are more stable, so the miRNA binding and degradation of the target genes may be more probable.

The AUGC frequency distribution at each of the nucleotide positions was analyzed for all miRNAs and compared with that for the stress regulated miRNAs by plotting line graphs from the data calculated from these sequences (**Figure 1**). When these graphs were compared for the normal and stress regulated sequences, the stress graph was seen to have a considerable variation (degree of randomness) at each position and for each base pair (A,U,G,C) thus proving that under stress, cell responds in a different way which can be studied further to understand its importance.

**Trend Analysis:**
The trend analysis graphs were plotted for normal and stress sequences to visualize the trend of frequency distribution for each nucleotide base. For nucleotide “A” the trend analysis showed to some extent a concave downward curve for stress regulated sequences, whereas a marginal linear increase graph for normal sequences (**Figure 2a**). The trend for nucleotide “U” showed a concave upward graph for stress regulated sequences, whereas a marginal linear decrease graph for normal sequences. Also the curve for stress lies below the trend line graph corresponding to normal sequences (**Figure 2b**).

Nucleotide “G” showed an altogether different trend of a significant concave downward curve for stress regulated sequences, indicating higher frequencies of this nucleotide in the mid positions of these sequences. The trend of normal sequences for this nucleotide showed a more or less constant graph (**Figure 2c**). For nucleotide “C” the trend of stress regulated sequences was opposite to that for G, as it showed a significant concave upward curve for stress regulated sequences, indicating higher frequencies of this nucleotide at the two ends of the sequences. Also curve for stress lies above the trend line graph corresponding to normal sequences. A slight linear increase graph was observed for the normal sequences (**Figure 2d**). These trends noticeably indicate that since the frequency of nucleotide “G” is higher in the mid positions and that of nucleotide “C” is higher at the two ends of the stress regulated sequences, the average GC content becomes high altogether for these sequences.

**Conclusion:**
Our results suggest that the GC content may play a important role in the interaction of miRNAs with their target genes. The significance of the degree of randomness of nucleotide frequency distribution in stress regulated miRNAs can be identified by future studies. An overall increase in the GC content of the stress regulated miRNAs further extends the view that high GC can be a critical parameter for prediction of stress regulated miRNAs in *Arabidopsis thaliana*.

**References:**

**[1]** SS Ambika *et al., Bioinformation* **2**: 431 (2008) [PMID: 18841238]
**[2]** R Sunkar *et al., Plant Cell* **16**:2001 (2004) [PMID: 15258262]
**[3]** B Zhang *et al., Comput Biol Chem.* **30**:395 (2006) [PMID: 17123865]
**[4]** B Zhang *et al., Gene* **397**:26 (2007) [PMID: 17574351]
**[5]** HH Kavi *et al., Bioessays* **27**:1209 (2005) [PMID: 16299769]
**[6]** PD Zamore *Nat Struct Biol.* **8**:746 (2001) [PMID: 11524674]
**[7]** GD John *et al., Genes Dev.* **17**: 438 (2003) [PMID: 12600936]
**[8]** C Llave *et al., Science* **297**:2053 (2002) [PMID: 12242443]
**[9]** T Hua *et al., Am J Hum Genet.* **76**: 268 (2005) [PMID: 15625622]
**[10]** P Dunoyer, O Voinnet *Curr Opin Plant Biol.* **8**:415 (2005) [PMID: 15939663]
**[11]** L Yan-du *et al., Plant Cell Rep.* **27**:1571 (2008) [PMID: 18626646]
**[12]** Y Bennasser *et al., Immunity* **22**:607 (2005) [PMID: 15894278]
**[13]** Y Jiao *et al., Plant Cell* **17**:1641 (2005) [PMID: 15863518]
**[14]** L Han-Hua et al., *RNA* **14**: 836 (2008) [PMID: 18356539]
**[15]** http://microrna.sanger.ac.uk/
**[16]** M Keisuke *et al., Nucleic Acids Res.* **37**: 1438 (2009) [PMID: 19136462]

**Edited by P. Kangueane**
**Citation: Akash *et al.*,** Bioinformation 4(4): 151-154 (2009)

## Supplementary material:

**Table 1:** Nucleotide frequency analysis of miRNAs

| Nucleotide base | All miRNAs | NaCl stress | Mannitol stress | 4°C stress | All three stresses |
|---|---|---|---|---|---|
| **A** | **26%** | **24%** | **24%** | **22%** | **22%** |
| **U** | **31%** | **24%** | **25%** | **27%** | **27%** |
| **G** | **23%** | **26%** | **26%** | **25%** | **26%** |
| **C** | **20%** | **26%** | **25%** | **26%** | **25%** |
| **GC CONTENT** | **43%** | **52%** | **51%** | **51%** | **51%** |

First group includes all the 123 miRNAs reported for *Arabidopsis thaliana*, NaCl stress corresponds to high salinity stress, Mannitol stress corresponds to drought stress, 4°C stress corresponds to cold stress and all three stress group corresponds to sequences which are regulated by all the three stresses.

**Table 2:** Dataset of 123 miRNAs in Arabidopsis thaliana obtained from miRBase

ath-miR156
ath-miR157
ath-miR158
ath-miR159
ath-miR160
ath-miR161
ath-miR162
ath-miR163
ath-miR164
ath-miR165
ath-miR166
ath-miR167
ath-miR168
ath-miR169
ath-miR170
ath-miR171
ath-miR172
ath-miR173
ath-miR319
ath-miR390
ath-miR391
ath-miR393
ath-miR394
ath-miR395
ath-miR396
ath-miR397
ath-miR398
ath-miR399
ath-miR400
ath-miR401
ath-miR402
ath-miR403
ath-miR404
ath-miR405
ath-miR406
ath-miR408
ath-miR413
ath-miR414
ath-miR415
ath-miR416
ath-miR417
ath-miR418
ath-miR419
ath-miR420
ath-miR426
ath-miR447
ath-miR771
ath-miR472
ath-miR773
ath-miR774
ath-miR775
ath-miR776
ath-miR777
ath-miR778
ath-miR779.1
ath-miR779.2
ath-miR780
ath-miR781
ath-miR782
ath-miR783
ath-miR822
ath-miR823
ath-miR824
ath-miR825
ath-miR826
ath-miR827
ath-miR828
ath-miR829.1
ath-miR829.2
ath-miR830
ath-miR831
ath-miR832-5p
ath-miR832-3p
ath-miR833-5p
ath-miR833-3p
ath-miR834
ath-miR835-5p
ath-miR835-3p
ath-miR836
ath-miR838
ath-miR839
ath-miR840
ath-miR841
ath-miR842
ath-miR843
ath-miR844
ath-miR845
ath-miR846
ath-miR847
ath-miR848
ath-miR849
ath-miR850
ath-miR851-5p
ath-miR851-3p
ath-miR852
ath-miR853
ath-miR854
ath-miR855
ath-miR856
ath-miR857
ath-miR858
ath-miR859
ath-miR860
ath-miR861-5p
ath-miR861-3p
ath-miR862-5p
ath-miR862-3p
ath-miR863-5p
ath-miR863-3p
ath-miR864-5p
ath-miR864-3p
ath-miR865-5p
ath-miR865-3p
ath-miR866-5p
ath-miR866-3p
ath-miR867
ath-miR868
ath-miR869.1
ath-miR869.2
ath-miR870
ath-miR1886
ath-miR1887
ath-miR1888