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; Vibha 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 sequence s 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

Methodology:
The miRNA 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.
GC content of stress induced sequences calculated by our method is 19.2% in Arabidopsis thaliana pairs. The average GC content of all miRNAs reported for Arabidopsis thaliana genome has been reported as just 36% due to the large number of miRNAs regulated by the three stresses. The average GC content of Arabidopsis thaliana sequences and all miRNAs (Table 1 in supplementary material) were added to these line graphs to compare the trend of nucleotide frequencies in a group of sequences calculated by the following formula:

\[ \text{Percentage content of N} = 100 \times \frac{\text{Sum of absolute frequencies of N}}{\text{Sum of total no. of sequences at each position}} \]

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 nucleotide frequencies in a group of sequences. The trend of nucleotide frequency distribution can be visualized using line graphs for each nucleotide base at 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 number of intronic sequences present within the genome. The average GC content of all miRNAs 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 at 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 is also a feature of genes coding for transcriptional activators or repressors which regulate gene expression under various stress conditions. 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 an 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] Y 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: 12600956]
[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: 18626464]
[12] Y Bennasser et al., Immunity 22:607 (2005) [PMID: 15904278]
[13] Y Jiao et al., Plant Cell 17:1641 (2005) [PMID: 15863518]
[14] L Han-Hua et al., RNA 14: 836 (2008) [PMID: 18356339]
[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)

License statement: This is an open-access article, which permits unrestricted use, distribution, and reproduction in any medium, for non-commercial purposes, provided the original author and source are credit.
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-miR397 | ath-miR775 | ath-miR834 | ath-miR858 |
|------------|------------|------------|------------|------------|
| ath-miR157 | ath-miR398 | ath-miR776 | ath-miR835-5p | ath-miR859 |
| ath-miR158 | ath-miR399 | ath-miR777 | ath-miR835-3p | ath-miR860 |
| ath-miR159 | ath-miR400 | ath-miR778 | ath-miR836 | ath-miR861-5p |
| ath-miR160 | ath-miR401 | ath-miR779.1 | ath-miR838 | ath-miR861-3p |
| ath-miR161 | ath-miR402 | ath-miR779.2 | ath-miR839 | ath-miR862-5p |
| ath-miR162 | ath-miR403 | ath-miR780 | ath-miR840 | ath-miR862-3p |
| ath-miR163 | ath-miR404 | ath-miR781 | ath-miR841 | ath-miR863-5p |
| ath-miR164 | ath-miR405 | ath-miR782 | ath-miR842 | ath-miR863-3p |
| ath-miR165 | ath-miR406 | ath-miR783 | ath-miR843 | ath-miR864-5p |
| ath-miR166 | ath-miR408 | ath-miR822 | ath-miR844 | ath-miR864-3p |
| ath-miR167 | ath-miR413 | ath-miR823 | ath-miR845 | ath-miR865-5p |
| ath-miR168 | ath-miR414 | ath-miR824 | ath-miR846 | ath-miR865-3p |
| ath-miR169 | ath-miR415 | ath-miR825 | ath-miR847 | ath-miR866-5p |
| ath-miR170 | ath-miR416 | ath-miR826 | ath-miR848 | ath-miR866-3p |
| ath-miR171 | ath-miR417 | ath-miR827 | ath-miR849 | ath-miR867 |
| ath-miR172 | ath-miR418 | ath-miR828 | ath-miR850 | ath-miR868 |
| ath-miR173 | ath-miR419 | ath-miR829.1 | ath-miR851-5p | ath-miR869.1 |
| ath-miR319 | ath-miR420 | ath-miR829.2 | ath-miR851-3p | ath-miR869.2 |
| ath-miR390 | ath-miR426 | ath-miR830 | ath-miR852 | ath-miR870 |
| ath-miR391 | ath-miR447 | ath-miR831 | ath-miR853 | ath-miR886 |
| ath-miR393 | ath-miR771 | ath-miR832-5p | ath-miR854 | ath-miR887 |
| ath-miR394 | ath-miR772 | ath-miR832-3p | ath-miR855 | ath-miR888 |
| ath-miR395 | ath-miR773 | ath-miR833-5p | ath-miR856 |
| ath-miR396 | ath-miR774 | ath-miR833-3p | ath-miR857 |