Send Orders for Reprints to reprints@benthamscience.net

Current Genomics, **2021**, *22*, 353-362

RESEARCH ARTICLE

In Silico Analysis of CCGAC and CATGTG *Cis-regulatory* Elements Across Genomes Reveals their Roles in Gene Regulation under Stress

Sneha Lata Bhadouriya[1], Abhishek Suresh[2], Himanshu Gupta[2], Sandhya Mehrotra[1], Divya Gupta[3] and Rajesh Mehrotra[1,*]

[1]Department of Biological Sciences, Birla Institute of Technology and Sciences Pilani, Goa, India; [2]Department of Biological Sciences, Birla Institute of Technology and Sciences Pilani, Pilani, India; [3]Shri Ramswaroop Memorial University, Lucknow, India

ARTICLE HISTORY

Received: February 15, 2021
Revised: April 01, 2021
Accepted: April 21, 2021

DOI:
10.2174/1389202922666210601095700

Abstract: ***Background*:** Plant yield closely depends on its environment and is negatively affected by abiotic stress conditions like drought, salinity, heat, and cold. Analysis of the stress-inducible genes in *Arabidopsis* has previously shown that CCGAC and CATGTG play a crucial role in controlling the gene expression through the binding of DREB/CBF and NAC TFs under various stress conditions, mainly drought and salinity.

***Methods*:** The pattern of these motifs is conserved, which has been analyzed in this study to find the mechanism of gene expression through spacer specificity, inter motif distance preference, functional analysis, and statistical analysis for four different plants, namely *Oryza sativa, Triticum aestivum, Arabidopsis thaliana,* and *Glycine max*.

***Results*:** The spacer frequency analysis has shown a preference for particular spacer lengths among four genomes. The spacer specificity at all the spacer lengths which predicts dominance of particular base pairs over others, was analyzed to find the preference of the sequences in the flanking region. Functional analysis on stress-regulated genes for saline, osmotic, and heat stress clearly shows that these motif frequencies with inter motif distance (0-30) in the promoter region of *Arabidopsis* are highest in genes which are upregulated by saline and osmotic stress and downregulated by heat stress.

***Conclusion*:** Microarray data were analyzed to confirm the role of both motifs in stress response pathways. Transcription factors seem to prefer larger motif size with repeated CCGAC and CATGTG elements. The common preference for one spacer was further validated through Box and Whisker's statistical analysis.

Keywords: Abiotic stress, *cis*-element, motif, spacer length, transcription factors, microarray.

1. INTRODUCTION

Gene expression and regulation in the eukaryotic organism have always been the topic for significant research. Abiotic stresses lead to a reduction in crop yield in various parts of the world. With the rising population, there comes the need for stress responsible plants with optimized gene-regulating promoters [1]. Transcriptional regulation is important in the regulation of stress-inducible genes. The binding of transcription factors to cis-regulatory elements is crucial for gene regulation. Various factors like promoters, enhancers, repressors, silencers play a major role in the binding of TFs to cis elements [2]. Alteration in Chromatin architecture helps the plants to respond during stress [3]. Cis-regulatory elements, which are the short-conserved region in the DNA present singly or in combination with other elements, are found to reoccur [4]. Together, they up-regulate or down-regulate the gene expression in response to stress. The transcription factor binding usually takes place with cis-elements with inter motif distance from 0 to 25 and is the sequence as well as location-specific. The role of different *cis*-regulatory elements in abiotic stresses in plants has been studied (W-box element under heat and salinity [5], the presence of high frequency of AAAGN7CTTT motif in *Arabidopsis thaliana* [6]). The present study will focus on spacer sequence analysis of two such motifs, CCGAC and CATGTG, in two dicots (*Arabidopsis thaliana* and *Glycine max*) and two monocots (*Oryza sativa* and *Triticum aestivum*). Arabidopsis has two pathways (i) abscisic acid (ABA) dependent and (ii) ABA independent pathways for gene expression under abiotic stress [1, 7]. Dehydration-responsive element/C-repeat (DRE/CRT) is involved in ABA independent regulatory pathway.

In *Arabidopsi*s, the over-expression of the DRE/CRT binding protein DREB1/CBF results in altered expression of

*Address correspondence to this author at the Department of Biological Sciences, Birla Institute of Technology and Sciences Pilani, Goa, India;
E-mail: rajeshm@goa.bits-pilani.ac.in

1389-2029/21 $65.00+.00 © 2021 Bentham Science Publishers

more than 40 stress-inducible genes, causing increase tolerance of freezing, salt, and drought conditions [8-10]. The transcriptional regulators that play a role in ABA-dependent regulatory systems include the MYC and MYB proteins [11]. In *Arabidopsis thaliana*, cis-elements and corresponding binding proteins, contain a distinct type of DNA binding domain, like AP2/ERF, HD-ZIP, basic leucine zipper, MYB, MYC, and several classes of zinc finger domains have been involved in plant stress responses as under several stress conditions, their expression is induced or repressed [1, 12].

In *Arabidopsis*, the expression of the RD29A/-COR78/LTI78 gene is up-regulated by cold, drought, and ABA [13]. The expression of this gene is found to be induced in both aba or *abi* mutants by drought and cold stresses, suggesting that under drought and cold stress conditions, both ABA-dependent and ABA-independent regulation are involved in it. The RD29A promoter was analyzed by deletion and base-substitution analyses. A 9-bp conserved sequence (TACCGACAT), known as the dehydration responsive element (DRE), is present in the RD29A promoter, which regulates RD29A expression in the ABA-independent response to dehydration and cold [14]. DRE does not require any other elements for its function as a single copy of DRE is sufficient for ABA-independent stress-responsive gene expression. DRE is also present in the promoter regions of many drought- and cold-inducible genes. Similarly, the A/GCCGAC motif is found in other cis-acting elements, like C-repeat (CRT) and low-temperature-responsive element (LTRE), showing regulation under cold stress. C-repeat-binding factor 1(CBF1), (DRE-binding protein 1A (DREB1A), and DREB2A are three cDNA encoding DRE isolated by the yeast one-hybrid screening method. These proteins particularly interact with the DRE/CRT sequence activating the gene transcription driven by the DRE/CRT sequence in *Arabidopsis*. CBF1 is similar to DREB1B, and its homologs, CBF2, and CBF3 are similar to DREB1C and DREB1A. DREB1/CBF gene expression is induced under cold stress and the expression of the DREB2 genes is induced under drought and high-salt stresses. Both DREB1 / CBF and DREB2 proteins bind to DRE, but DREB1 / CBFs are believed to be involved in cold-responsive gene expression while DREB2 are involved in drought-responsive gene expression, suggesting indicating a cross-talk between drought-and cold-responsive gene expression occurs on a cis-acting element DRE [15].

CATGTG is the MYC-like sequence that plays a crucial role in dehydration and salinity-inducible expression of the EARLY RESPONSIVE TO DEHYDRATION STRESS 1 (ERD1) gene in *Arabidopsis thaliana* [16], which encodes a ClpA (ATP binding subunit of the caseinolytic ATP-dependent protease) homologous protein [17]. This core sequence is functionally and biologically important in many promoters that respond to stresses like dehydration, salinity, and abscisic acid [18]. This core motif occurs in multiple numbers at different spacer lengths upstream of many genes. Any change in this may affect the overall gene expression as the motif contributes directly by stabilizing the transcription site binding and the complex so formed. Three NAC transcription factors (ANAC019, ANAC055, and ANAC072/RD26) were found to bind the MYC-like sequence present in the promoter of ERD1 [19]. NAC transcription factors involved in stress response and tolerance have been classified in the stress-responsive NAC (SNAC) group. NAC transcription factors have a highly conserved N-terminal DNA-binding domain and a variable C-terminal region. It is believed that this C-terminal region plays a crucial role in determining its target genes [18].

NAC transcription factors are engaged with different developmental processes, from shoot meristem development to auxin signalling [20]. Some reports exhibit its participation in biotic and abiotic stress signalling [18, 20]. Detailed DNA binding assay of these NAC transcription factors decided NACRS (NAC recognition sequence) ANNNNNTCNNNNNNNACACGCATGT, which contains CATGT and harbours CACG as the core DNA binding site [19]. These NAC genes were found to be expressed within 1-2 h of ABA treatment, suggesting that they are induced through an ABA-independent pathway under drought stress [19]. The transgenic plants that overexpress ANAC019 and ANAC072 showed a phenotype and growth time course similar to the vector control. In contrast, the plants that overexpress ANAC055 showed a growth rate like that of the vector control until they reached the rosette stage; after this point, plants having the expression of the transgene at the medium level, indicated a little delay in bolting when compared to vector control whereas; plants in which the transgene was overexpressed at high level remained at rosette stage for an additional few days before first bolting. The transgenic plant that overexpressed ANAC019, ANAC055 or ANAC072 / RD26 induced the expression of many stress-inducible genes but was unable to up-regulate ERD1 [19, 21].

Following the sequence and inter motif distance specificity, our study attempts to analyze the spacer frequency data with the distance varying from N=0-30, across four plant genomes for CCGAC and CATGTG motif combinations. The frequency of these motifs was also compared with some control sequences which were identified not to be conserved themselves [22]. CCGAC and CATGTG occurrence with ACGT elements was analyzed to understand their role in stress regulation [23]. Consensus spacer frequency for each plant genome for CCGAC and CATGTG repeat motif combinations were also identified. Functional analysis of the *Arabidopsis* promoter gives the trend of the correlation spacer frequency and gene regulation under particular stress among salinity, osmotic, and heat. Further statistical analysis was performed through box and whiskers analysis of the spacer frequency of CCGAC and CATGTG repeat combinations, showing that some inter-motif distances are more prevalent in both dicot and monocot species. Transcription factor binding sites were also found using JASPAR and confirmed using CONSITE [24]. The knowledge of such *cis*-regulatory elements in plant stresses will further help understand the plant genome structure [25]. With the use of such information, synthetic modules can be created to increase crop yield [26].

2. METHODOLOGY

2.1. Data Extraction

Firstly, the genomic data for four species, two monocots - *Oryza sativa* (International Rice Genome Sequencing Project, Build 4.0, 2009) [27] and *Triticum aestivum*, two dicots - *Arabidopsis thaliana* (The *Arabidopsis* genome initiative v. 10, 2011) [28] and *Glycine max* (US DOE Joint Genome Institute (JGI-PGF) was obtained from NCBI database. Spacer frequency analysis was done on the genomic data to find the optimum spacer distance between co-occurring CCGAC and CATGTG elements. For this, spacer lengths of N=0-30 were analyzed for both the motif repeats. Transcription factors mainly bind within a distance of 25 base pairs and hence the limitation to N=30. To test the biological significance of these motifs, we generated four control sequences (ACGCC, CAGCC, AGCCC, and CCGCA) for CCGAC (GTAGCT, AGGCTT, GGTCTA, and TAGCTG) for CATGTG and performed a similar analysis. We ensured that none of the controls was cis-regulatory elements themselves using the PLANT CARE database [22]. We, then, compared the frequency of occurrence of the CCGAC(N)CCGAC and CATGTG(N)CATGTG motifs with that of different control sequences for N=0-30.

2.2. Spacer Sequence Analysis

Spacer sequence in the overall genome was analyzed to check the specificity for particular base pairs in the flanking region of these motifs for N=0-30. The genome-wide GC content was calculated. To identify the nucleotide preference, a threshold of 70% was taken *i.e.*, if the frequency of G or C was more than 70% of the GC-content, we assumed that G or C is the preferred base. Similarly, if A or T had a frequency of occurrence of more than 70% of (1-GC content) at that particular position, then A or T is the preferred base. For *e.g.* in the case of *Arabidopsis* the total Content in the genome was 36%, hence to normalize the threshold for G and C was taken to be 25% and A and T were taken to be 40%. In this way, the consensus sequences for all spacers were generated.

2.3. Functional Analysis

Using the *Arabidopsis thaliana* microarray data published from the EBI gene expression atlas [29], we analyzed if the up-regulated /down-regulated genes have CCGAC and CATGTG motifs suggesting their role in stress response. The stresses that were focussed here include heat, osmotic, and saline stress. To identify the correlation between the occurrence of these motifs with ACGT cis-element, the combinations with ACGT (CCGAC_ACGT and ACGT_CCGAC; CATGTG_ACGT and ACGT_CATGTG) were also analyzed.

2.4. Statistical Analysis

Box and Whiskers analysis was performed to check the position of the relation of the outliers with the boxes to check the important peaks which are common across the genome in an attempt to establish correlation among the four genomes. To test the statistical significance, a paired student t-test was conducted using the standard protocol [30]. For this, the frequency of occurrence from 0 to 30 between CCGAC(N)CCGAC and CATGTG(N)CATGTG motifs were compared with the control sequences and to test the statistical significance of the sequences, the paired t-test was applied. The t-test was also applied for the occurrence of combinations of these motifs with ACGT.

2.5. Transcription Factor Binding Site

The *Arabidopsis thaliana* promoter sequence was analyzed for CCGAC and CATGTG motifs and their spacer sequences. Following this, we found the transcription factors binding to these cis-elements separated by a few base pair distance. A 139 bp minimal promoter sequence was used for this study. TCACTATATATAGGAAGTTCATTTCATTTGGAATGGACACGTGTTGTCATTTCTCAACAATTACCAACAACAACAAACAACAAACAACATTATACAATTACTATTTACAATTACATCTAGATAAACAATGGCTTCCTCC The MPS was suffixed to a few of the spacer sequences. The above sequence along with its spacer sequence version was used in JASPAR to identify the transcription factors that would bind at various sites. The results were then confirmed using CONSITE [24, 31].

3. RESULTS

The frequency of occurrence of CCGAC and CATGTG was the maximum when compared to the control sequences in all four species. This suggests that the core motifs being analyzed are biologically more significant (Fig. **1**). The spacer analysis showed that the occurrence of CCGAC_CCGAC and CATGTG_CATGTG separated by varying spacer distance has a much larger frequency when compared to the different control sequences (Fig. **2**). To test the significance of the results, a student t-test was conducted which showed that the frequencies of occurrence are statistically significant ($p<0.05$ for CCGAC_CCGAC in both dicots and monocots while $p<0.001$ for CATGTG_CATGTG in dicots and monocots $p<0.05$). The overall frequency of CCGAC_CCGAC was found to be 998 in *Arabidopsis thaliana*, 4670 in *Glycine max*, 12608 in *Oryza sativa,* and 465494 in *Triticum aestivum*. In the case of CATGTG_CATGTG the overall frequency was calculated to be 415 in *Arabidopsis thaliana*, 6733 in *Glycine max*, 2912 in *Oryza sativa*, and 109497 in *Triticum aestivum*. On analysis, our data showed the variable frequency for N=0-30 across the four species. In the case of CATGTG_CATGTG, potential common peaks were identified at N= 9,27 in dicots and N=2,11 in monocots (Fig. **3**) high degree of overlapping/ correlation was found in the case of monocots for N = 8 to 12. Whereas in the case of CCGAC_CCGAC, spacer frequency data shows common dips and peaks across the four plant genomes, particularly at the motif length 1 except *Arabidopsis thaliana.* Spacer frequency analysis also showed a high correlation among monocots for spacer length 0-10. Common dips were observed among the four-plant species except *Arabidopsis thaliana* at inter motif distances of 6 and 18 which further indicates the

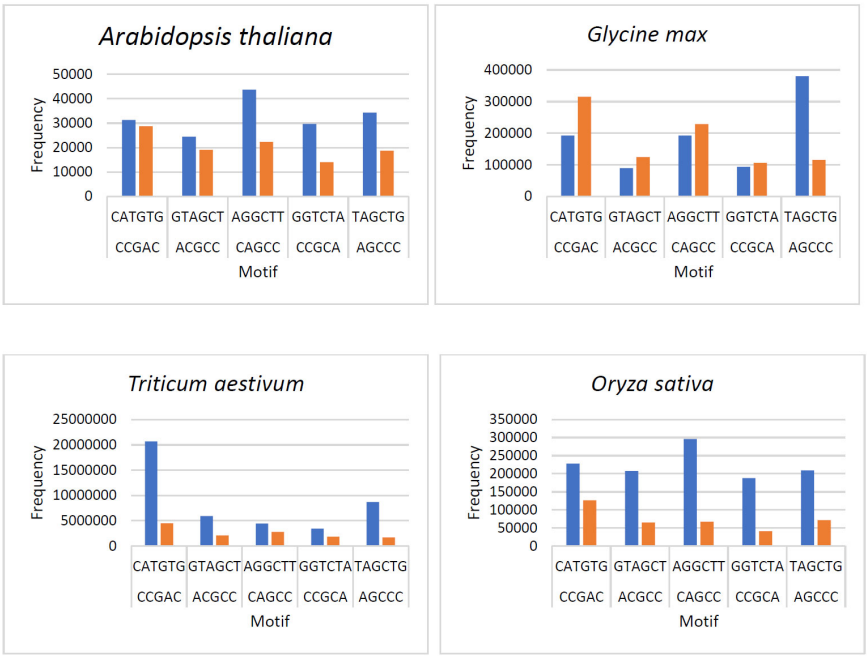

Fig. (1). The frequency of occurrence of CCGAC and CATGTG compared to the control sequences among four plant species. The highest frequency of the CCGAC and CATGTG suggests that these are biologically significant. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

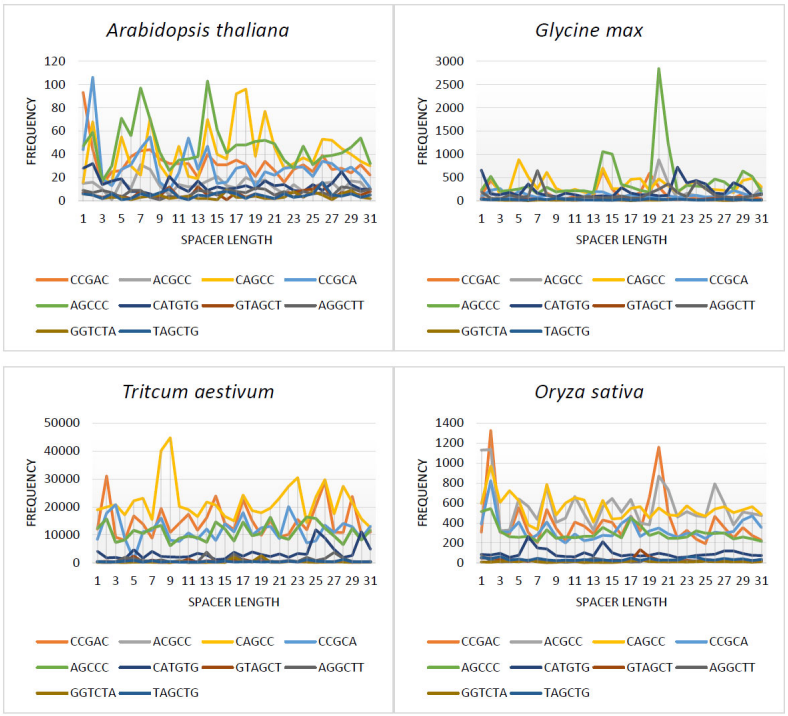

Fig. (2). The spacer analysis and the occurrence of CCGAC_CCGAC, CATGTG_CATGTG separated by varying spacer distance compared to the different control sequences in all the four plant genomes. It shows that the occurrence of CCGAC_CCGAC and CATGTG_CATGTG separated by varying spacer distance has a much larger frequency in comparison to the different control sequences. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

deviation of *Arabidopsis thaliana* frequency pattern from the rest of the genomes (Fig. **3**). Following this, a box and whisker plot was made to get a better understanding of the varying frequency of occurrence (Fig. **4**).

Comparison with controls and combinations of DRE and CATGTG with ACGT gives little inference that the ACGT repeat motif has the highest frequency among the four possible combinations. The student t-test was applied for identifying the statistical significance ($p<0.001$, $t=4.152$). Potential common peaks were identified at N = 1,6 for CATGTG_ACGT in monocots. No common peaks were found for dicots. Similarly, potential common peaks identified for ACGT_CATGTG include N=1, 3, 5, 7 in monocots and N=1 for dicots. A high degree of overlapping was seen for ACGT_CATGTG in monocots for N = 0 to 2. CCGAC was comparably preferable among the four species, with wheat having the highest frequency of CCGAC. CAGCC control was especially afvoured in terms of spacer frequency across the four species (Fig. **5**). Consensus sequence analysis shows conserved terminal and initial positions within a spacer sequence and it was found that G is more prevalent in *Arabidopsis thaliana* and wheat at the first position after the motif in the spacer. There was limited prevalence for other regions in the spacers for all of the four species (Fig. **6**). While in the case of CATGTG_CATGTG, there is no specific preference for any base at the starting and terminal position the case of. However, it was seen that G was the preferred base in the first position when ACGT was the first motif. Similarly, C was the preferred base at the terminal position when ACGT was the second motif. The consensus sequences for CATGTG_CATGTG and CATGTG with ACGT have more than one conserved sequence compared to that of ACGT_ACGT spacer 24 which was the only sequence conserved (Fig. **6**). Functional analysis on stress-regulated genes for saline, osmotic, and heat stress clearly shows that CCGAC motif frequency with inter motif distance (0-30) in the promoter region of *Arabidopsis* is highest in genes that are upregulated by osmotic stress and downregulated by heat stress.

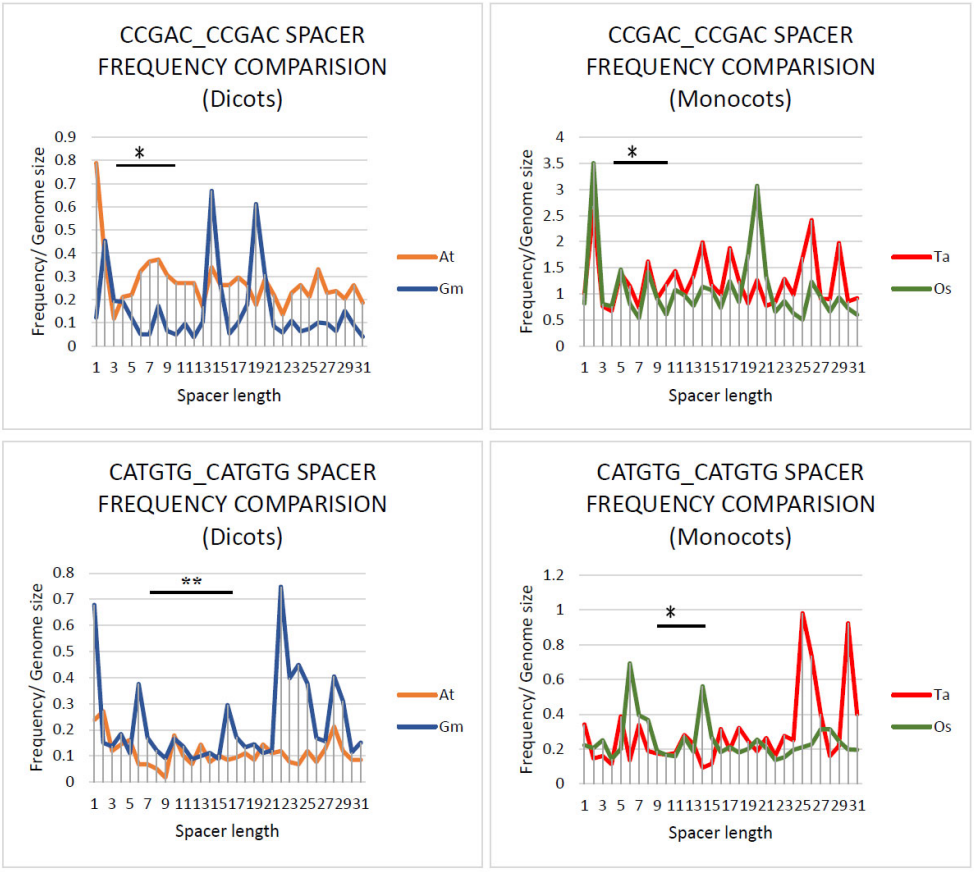

Fig. (3). (**a**) Spacer frequency comparison of CCGAC_CCGAC in monocots and dicots. CCGAC_CCGAC, spacer frequency data shows common dips and peaks across the four plant genomes, particularly at the motif length 1 except *Arabidopsis thaliana.* (**b**) Spacer frequency comparison of CATGTG_CATGTG in monocots and dicots. In the case of CATGTG_CATGTG, potential common peaks were identified at N= 9,27 in dicots and N=2,11 in monocots, high degree of overlapping/ correlation was found in the case of monocots for N = 8 to 12. (At: *Arabidopsis thaliana*; Gm: *Glycine max*; Ta: *Triticum aestivum*; Os: *Oryza sativa*). * $p< 0.05$; ** $p< 0.001$. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

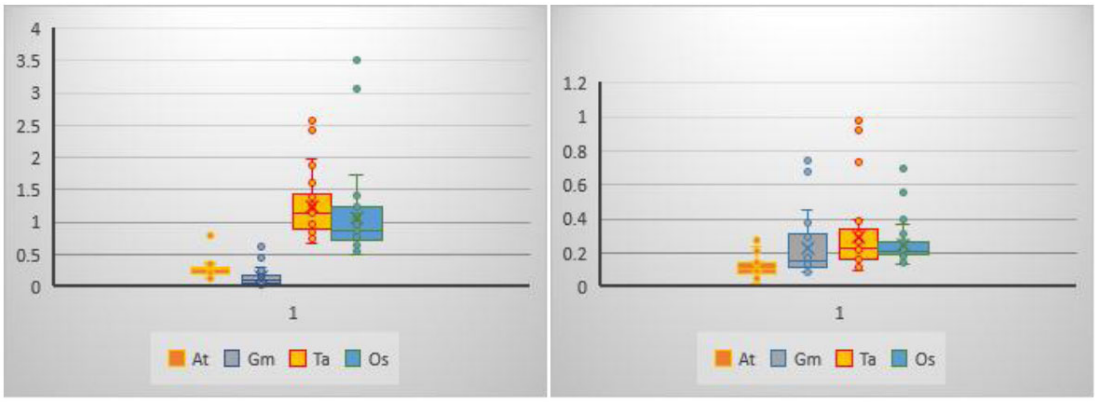

Fig. (4). Box and whisker plot to get a better understanding of the varying frequency of occurrence of CCGAC_CCGAC and CATGTG_-CATGTG among four plant genomes. (At: *Arabidopsis thaliana*; Gm: *Glycine max*; Ta: *Triticum aestivum*; Os: *Oryza sativa*). (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

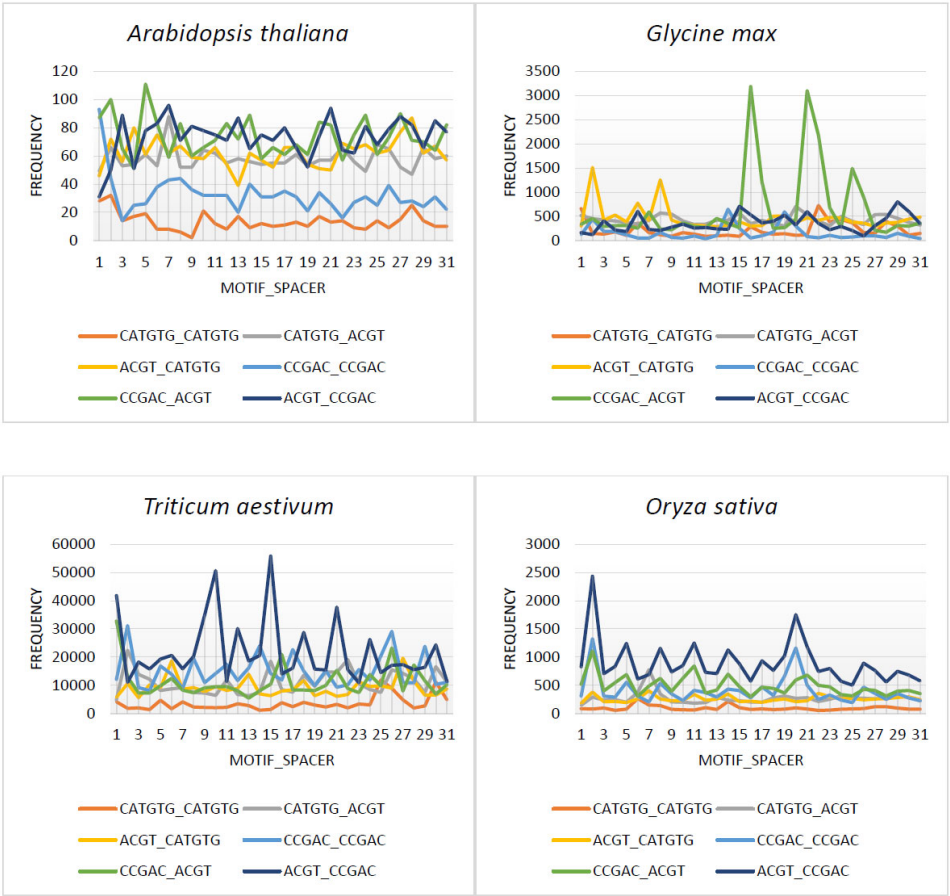

Fig. (5). Comparison with controls and combinations of CCGAC and CATGTG with ACGT among four plant genomes. It shows that the ACGT repeat motif has the highest frequency among the four possible combinations. Potential common peaks were identified at N = 1,6 for CATGTG_ACGT in monocots. No common peaks were found for dicots. Similarly, potential common peaks identified for ACGT_CAT-GTG include N=1, 3, 5, 7 in monocots and N=1 for dicots. A high degree of overlapping was seen for ACGT_CATGTG in monocots for N = 0 to 2. CCGAC was comparably preferable among the four species, with wheat having the highest frequency of CCGAC CCGAC_CCGAC (In *Arabidopsis thaliana*) CATGTG_CATGTG (In *Arabidopsis thaliana*). (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

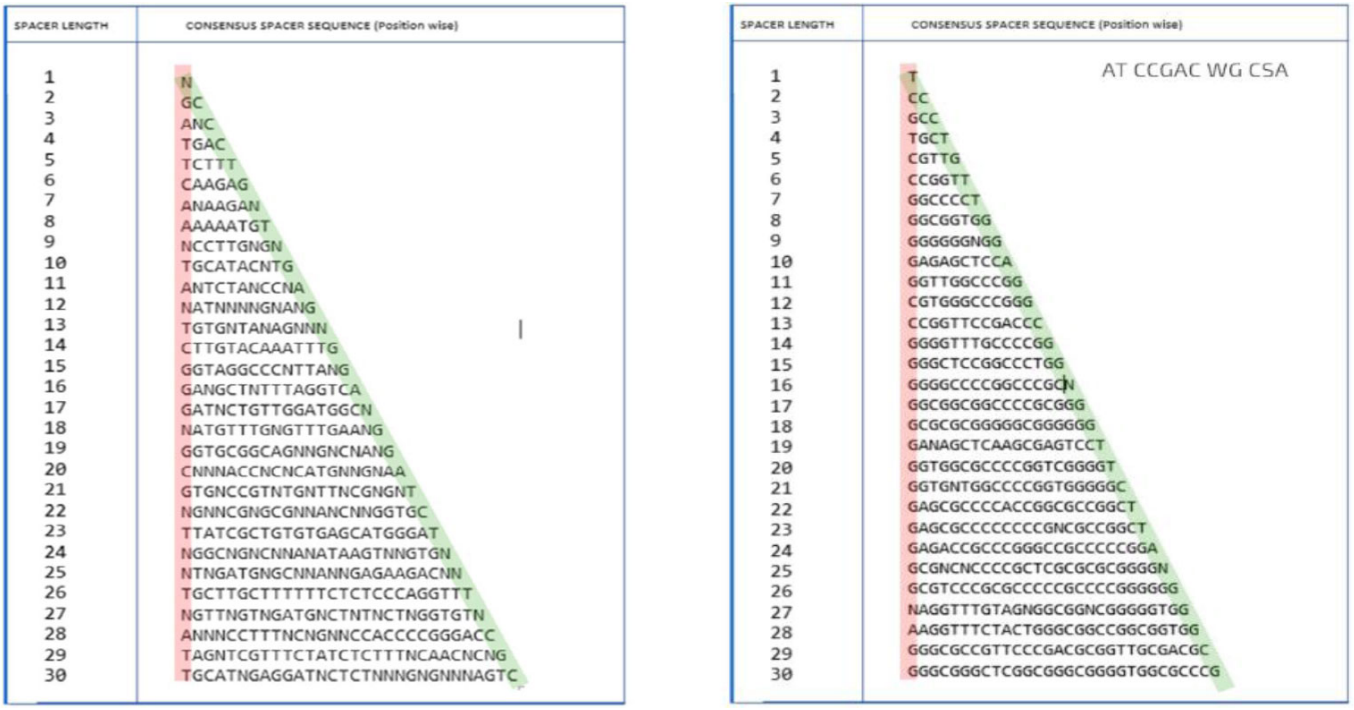

Fig. (6). Consensus sequence analysis of CCGAC_CCGAC and CATGTG_CATGTG in *Arabidopsis thaliana*. It shows conserved terminal and initial positions within a spacer sequence and it was found that G is more prevalent in *Arabidopsis thaliana* the first position after the motif in the spacer. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

Microarray data that were analyzed showed that the motif CATGTG plays a crucial role in gene up/downregulation in response to different stress conditions. From the data obtained, it was seen that the frequency of CATGTG_CATGTG is more in the case of genes down-regulated by heat, upregulated by saline and osmotic stress. Comparing the results for different stresses shows that CATGTG along with ACGT plays a major role in genes upregulated by heat. And among the downregulated genes, it was seen that the CATGTG_CATGTG does not follow the same trend as its ACGT combinations (Fig. **7**). CCGAC repeat and combinations with ACGT are regulated by osmotic and heat stress which was evident after analyzing the spacer frequencies in the genes upregulated and downregulated by heat, salinity, and osmotic stresses in the *Arabidopsis* promoter region. It was found the DRE plays an important role in the upregulation of osmotic stress-inducible genes and downregulation of heat stress-inducible genes. Further, the functional analysis also showed that DRE in combination with ACGT plays a major role in the down-regulation of heat-inducible genes and upregulation of osmotic stress-inducible genes (Fig. **7**).

The minimal promoter sequence along with the spacer sequence was analyzed for the transcription factor binding site using the JASPAR database. It was seen that by adding the spacer sequence before the MPS, the no. of putative sites predicted increased. Increasing the spacing between the motifs seemed to increase the number of transcription factor binding sites in most of the cases. Tandemly repeating CCGAC and CATGTG motif had the maximum number of transcription sites binding to them. MPS alone has 27 transcription factor binding sites. Adding the CCGAC and CATGTG motif, the putative sites increase to 28 and 29, respectively. Increasing the copy number of the motif *i.e.*, repeating it tandemly twice or thrice shows a higher number of binding sites (36 and 42 respectively in case of CCGAC) (40 and 49 respectively in case of CATGTG). CATGTG motif itself is a binding site to the MYC transcription factor. The result was further confirmed using CONSITE.

4. DISCUSSION

The main factors affecting gene expression and regulation include the interaction of transcription factors and the cis-elements. NAC and DREB/CBF transcription factors are part of the largest families of transcription regulators in plants. They play a major role in abiotic stress response [18, 32]. TaNAC29 a transcription factor from wheat has been shown to enhance salt and drought tolerance in *Arabidopsis thaliana* [33]. In all the cases with involvement to NAC transcription factor, it was seen that the protein involved bound to the MYC recognition sequence CATGTG [19]. So we started by analyzing the spacer sequences for this particular motif. DRE is involved in ABA independent regulatory pathway and involves regulating gene expression under abiotic stress.

For the analysis, we considered four control sequences (ACGCC, CAGCC, AGCCC, and CCGCA) for CCGAC and (GTAGCT, AGGCTT, GGTCTA, and TAGCTG) for

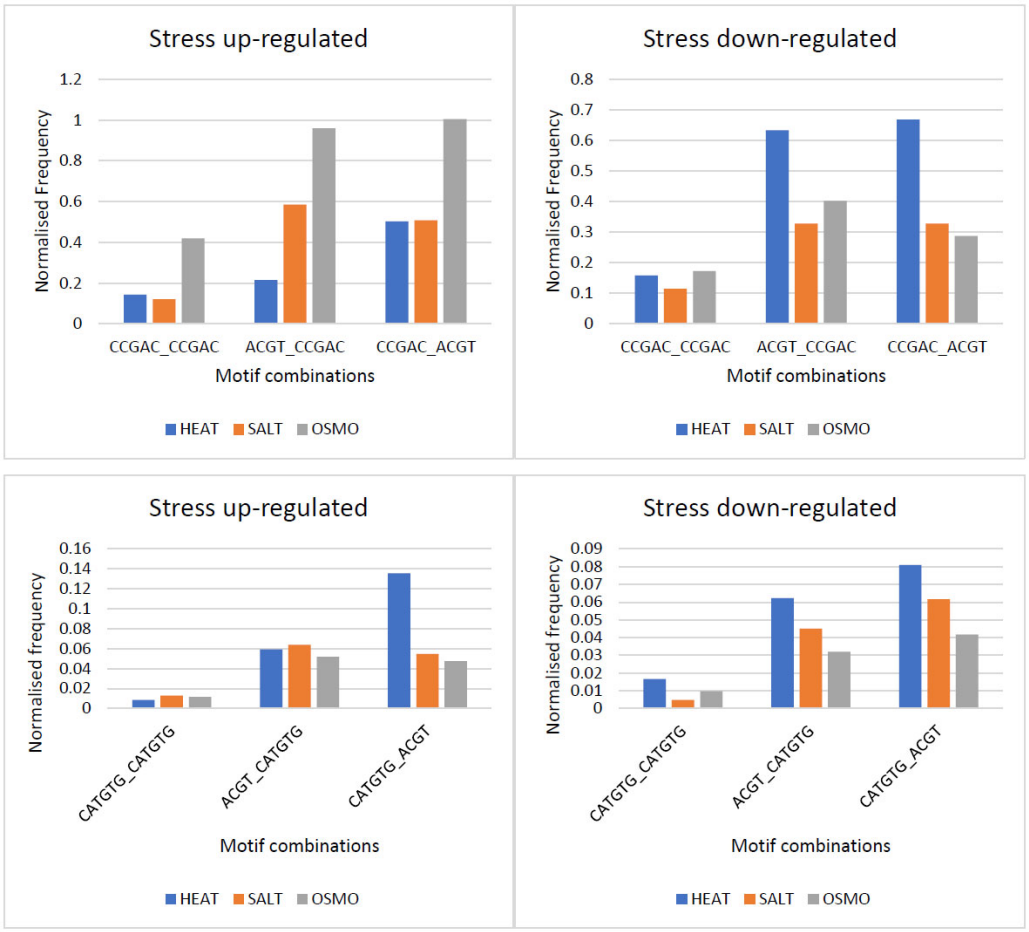

Fig. (7). Functional role of CATGTG and CCGAC under stress conditions in four plant genomes. It shows that these motifs play a crucial role in gene up/downregulation in response to different stress conditions. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

CATGTG. PLANT CARE database was used to ensure that none of the control sequences chosen were cis-elements [22]. It was seen that the individual frequency across the genome was maximum in the case of CCGAC and CATGTG suggesting their biological significance. The overall spacer frequency for CCGAC_CCGAC and CATGTG_CATGTG was the maximum compared to the control sequences. In the case of CCGAC_CCGAC, common peaks were found in monocots for N= 1-10 and dicots at N=1. *Arabidopsis* was observed showing deviation in the spacer frequency pattern among four species. In the case of CATGTG_CATGTG common peaks were found in dicots for N= 9, 27 and monocots for N=2,11. A high degree of overlapping/ correlation was found in the case of monocots for N = 8 to 12 suggesting it to be the local region of conservation. The result was found statistically significant using a student t-test.

The spacer frequency observed for our motif in combination with ACGT had a higher frequency than CCGAC_CCGAC and CATGTG_CATGTG. This is probably because of the involvement of the ACGT element in various stress responses [3]. Potential common peaks were identified at N = 1,6 for CATGTG_ACGT in monocots. No common peaks were found for dicots. Similarly, potential common peaks identified for ACGT_CATGTG include N=1, 3, 5, 7 in monocots and N=1 for dicots. A high degree of overlapping was seen for ACGT_CATGTG in monocots for N = 0 to 2, suggesting it to be the local region of conservation.

The consensus sequence showed a clear preference for G at the first position with ACGT as the first motif and C at the terminal position with ACGT as the second motif. The CACGTG motif also recognized as G-box has known to enhance the Transcription factor binding to ACGT elements [34].

The microarray data analyzed show that the presence of CCGAC and CATGTG motifs in various genes up-regulated and down-regulated by various stresses suggesting the involvement of these motifs in the respective response pathways. It was seen that the motifs are mainly involved with genes down-regulated by heat, up-regulated by saline and osmotic stress. However, to the contrary, it was seen that the frequency of occurrence of our motif in combination with

ACGT was the maximum in the case of genes up-regulated by heat. This result is quite peculiar when we look at the gene regulation/expression affected by heat alone. In the case of genes down-regulated by various stress factors, it was seen the ACGT combinations do not follow a similar trend as the CATGTG_CATGTG itself. The ACGT combinations violate the trend in the case of salt stress. This implies that ACGT is closely involved with CATGTG when it comes to gene expression regulated by salt stress. Statistical analysis through box and whiskers plot gives a clear image of the pattern of spacer outliers. It can be seen that inter motif distance one frequency is an outlier in the median box region across the five species. Common peaks at one (CCGAC_N_CCGAC and CATGTG_N_CATGTG) motif might have been conserved while evolution and plays an important role in heat, drought, and cold response.

CONCLUSION

DRE (Drought response element) and CATGTG are important cis-elements that play an important role in the regulation of cold, heat, and osmotic stress-inducible genes as a favourable transcription across four genomes. Spacer frequency analysis showed the correlation among the four species at a spacer length of one. Spacer sequence analysis showed that G was prevalent base pair in wheat and *Arabidopsis* in the first position after the start motif. It is also evident that ACGT plays an important role in combination with CCGAC and CATGTG in stress regulation. Functionally analyzing the *Arabidopsis* promoter makes it evident that CCGAC and CATGTG are involved in the downregulation of heat-inducible genes and upregulation of osmotic stress regulation genes. CATGTG cis-elements in their promoter region, implying their significant role in the respective pathways. Increased numbers of transcription factor binding sites with tandem CCGAC and CATGTG repeats show a preference/evolution of larger size motifs in *Arabidopsis thaliana* [35]. Our laboratory is interested in designer promoters and the results obtained from the present study have great application in such studies. In conclusion, from this study, more insight was covered through the analysis of genomes of four species, about the patterns and evolution of CCGAC and CATGTG cis-elements.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

Not applicable.

HUMAN AND ANIMAL RIGHTS

No animals/humans were used for studies that are the basis of this research.

CONSENT FOR PUBLICATION

Not applicable.

AVAILABILITY OF DATA AND MATERIALS

Not applicable.

FUNDING

This work was supported by SERB project EMR/2016/002470 sanctioned by the government of India to S.M. and R.M.

CONFLICT OF INTEREST

The authors declare no conflict of interest, financial or otherwise.

ACKNOWLEDGEMENTS

R.M. and S.M. are thankful to the Department of Science and Technology for providing financial support. This work was supported by SERB project EMR/2016/002470 sanctioned by the government of India to S.M. and R.M. We would like to thank the Birla Institute of Technology and Sciences, Pilani, India for providing support and fellowship to S.L.

REFERENCES

[1] Shinozaki, K.; Yamaguchi-Shinozaki, K. Molecular responses to dehydration and low temperature: Differences and cross-talk between two stress signaling pathways. *Curr. Opin. Plant Biol.,* **2000**, *3*(3), 217-223. http://dx.doi.org/10.1016/S1369-5266(00)80068-0 PMID: 10837265

[2] Khan, Z.; Kumar, B.; Dhatterwal, P.; Mehrotra, S.; Mehrotra, R. Transcriptional regulatory network of cis-regulatory elements (Cres) and transcription factors (Tfs) in plants during abiotic stress. *Int. J. Plant Biol. Res.,* **2017**, *5*(2), 1064-1081.

[3] Bhadouriya, S.L.; Mehrotra, S.; Basantani, M.K.; Loake, G.J.; Mehrotra, R. Role of chromatin architecture in plant stress responses: An update. *Front. Plant Sci.,* **2021**, *11*, 603380. http://dx.doi.org/10.3389/fpls.2020.603380 PMID: 33510748

[4] Mehrotra, R.; Sethi, S.; Zutshi, I.; Bhalothia, P.; Mehrotra, S. Patterns and evolution of ACGT repeat cis-element landscape across four plant genomes. *BMC Genomics,* **2013**, *14*, 203. http://dx.doi.org/10.1186/1471-2164-14-203 PMID: 23530833

[5] Dhatterwal, P.; Basu, S.; Mehrotra, S.; Mehrotra, R. Genome wide analysis of W-box element in *Arabidopsis thaliana* reveals TGAC motif with genes down regulated by heat and salinity. *Sci. Rep.,* **2019**, *9*(1), 1681. http://dx.doi.org/10.1038/s41598-019-38757-7 PMID: 30737427

[6] Mehrotra, R.; Jain, V.; Shekhar, C.; Mehrotra, S. Genome wide analysis of *Arabidopsis thaliana* reveals high frequency of AAAGN7CTTT motif. *Meta Gene,* **2014**, *2*(2), 606-615. http://dx.doi.org/10.1016/j.mgene.2014.05.003 PMID: 25606443

[7] Mehrotra, R.; Bhalothia, P.; Bansal, P.; Basantani, M.K.; Bharti, V.; Mehrotra, S. Abscisic acid and abiotic stress tolerance - different tiers of regulation. *J. Plant Physiol.,* **2014**, *171*(7), 486-496. http://dx.doi.org/10.1016/j.jplph.2013.12.007 PMID: 24655384

[8] Seki, M.; Narusaka, M.; Abe, H.; Kasuga, M.; Yamaguchi-Shinozaki, K.; Carninci, P.; Hayashizaki, Y.; Shinozaki, K. Monitoring the expression pattern of 1300 *Arabidopsis* genes under drought and cold stresses by using a full-length cDNA microarray. *Plant Cell,* **2001**, *13*(1), 61-72. http://dx.doi.org/10.1105/tpc.13.1.61 PMID: 11158529

[9] Fowler, S.; Thomashow, M.F. *Arabidopsis* transcriptome profiling indicates that multiple regulatory pathways are activated during cold acclimation in addition to the CBF cold response pathway. *Plant Cell,* **2002**, *14*(8), 1675-1690. http://dx.doi.org/10.1105/tpc.003483 PMID: 12172015

[10] Maruyama, K.; Sakuma, Y.; Kasuga, M.; Ito, Y.; Seki, M.; Goda, H.; Shimada, Y.; Yoshida, S.; Shinozaki, K.; Yamaguchi-Shinozaki, K. Identification of cold-inducible downstream genes of the *Arabidopsis* DREB1A/CBF3 transcriptional factor using two microarray systems. *Plant J.,* **2004**, *38*(6), 982-993.

http://dx.doi.org/10.1111/j.1365-313X.2004.02100.x PMID: 15165189

[11] Abe, H.; Urao, T.; Ito, T.; Seki, M.; Shinozaki, K.; Yamaguchi-Shinozaki, K. *Arabidopsis AtMYC2 (* bHLH) and *AtMYB2* (MYB) function as transcriptional activators in abscisic acid signaling. *Plant Cell,* **2003**, *15*(1), 63-78. http://dx.doi.org/10.1105/tpc.006130 PMID: 12509522

[12] Pastori, G.M.; Foyer, C.H. Common components, networks, and pathways of cross-tolerance to stress. The central role of "redox" and abscisic acid-mediated controls. *Plant Physiol.,* **2002**, *129*(2), 460-468. http://dx.doi.org/10.1104/pp.011021 PMID: 12068093

[13] Yamaguchi-Shinozaki, K.; Shinozaki, K. Organization of cis-acting regulatory elements in osmotic- and cold-stress-responsive promoters. *Trends Plant Sci.,* **2005**, *10*(2), 88-94. http://dx.doi.org/10.1016/j.tplants.2004.12.012 PMID: 15708346

[14] Yamaguchi-Shinozaki, K.; Shinozaki, K. A novel cis-acting element in an Arabidopsis gene is involved in responsiveness to drought, low-temperature, or high-salt stress. *Plant Cell,* **1994**, *6*(2), 251-264. http://dx.doi.org/10.2307/3869643 PMID: 8148648

[15] Ito, Y.; Katsura, K.; Maruyama, K.; Taji, T.; Kobayashi, M.; Seki, M.; Shinozaki, K.; Yamaguchi-Shinozaki, K. Functional analysis of rice DREB1/CBF-type transcription factors involved in cold-responsive gene expression in transgenic rice. *Plant Cell Physiol.,* **2006**, *47*(1), 141-153. http://dx.doi.org/10.1093/pcp/pci230 PMID: 16284406

[16] Simpson, S.D.; Nakashima, K.; Narusaka, Y.; Seki, M.; Shinozaki, K.; Yamaguchi-Shinozaki, K. Two different novel cis-acting elements of erd1, a clpA homologous *Arabidopsis* gene function in induction by dehydration stress and dark-induced senescence. *Plant J.,* **2003**, *33*(2), 259-270. http://dx.doi.org/10.1046/j.1365-313X.2003.01624.x PMID: 12535340

[17] Nakashima, K.; Kiyosue, T.; Yamaguchi-Shinozaki, K.; Shinozaki, K. A nuclear gene, erd1, encoding a chloroplast-targeted Clp protease regulatory subunit homolog is not only induced by water stress but also developmentally up-regulated during senescence in *Arabidopsis thaliana. Plant J.,* **1997**, *12*(4), 851-861. http://dx.doi.org/10.1046/j.1365-313X.1997.12040851.x PMID: 9375397

[18] Nuruzzaman, M.; Sharoni, A.M.; Kikuchi, S. Roles of NAC transcription factors in the regulation of biotic and abiotic stress responses in plants. *Front. Microbiol.,* **2013**, *4*, 248. http://dx.doi.org/10.3389/fmicb.2013.00248 PMID: 24058359

[19] Tran, L.S.; Nakashima, K.; Sakuma, Y.; Simpson, S.D.; Fujita, Y.; Maruyama, K.; Fujita, M.; Seki, M.; Shinozaki, K.; Yamaguchi-Shinozaki, K. Isolation and functional analysis of *Arabidopsis* stress-inducible NAC transcription factors that bind to a drought-responsive cis-element in the early responsive to dehydration stress 1 promoter. *Plant Cell,* **2004**, *16*(9), 2481-2498. http://dx.doi.org/10.1105/tpc.104.022699 PMID: 15319476

[20] Olsen, A.N.; Ernst, H.A.; Leggio, L.L.; Skriver, K. NAC transcription factors: structurally distinct, functionally diverse. *Trends Plant Sci.,* **2005**, *10*(2), 79-87. http://dx.doi.org/10.1016/j.tplants.2004.12.010 PMID: 15708345

[21] Fujita, M.; Fujita, Y.; Maruyama, K.; Seki, M.; Hiratsu, K.; Ohme-Takagi, M.; Tran, L.S.; Yamaguchi-Shinozaki, K.; Shinozaki, K. A dehydration-induced NAC protein, RD26, is involved in a novel ABA-dependent stress-signaling pathway. *Plant J.,* **2004**, *39*(6), 863-876. http://dx.doi.org/10.1111/j.1365-313X.2004.02171.x PMID: 15341629

[22] Lescot, M.; Déhais, P.; Thijs, G.; Marchal, K.; Moreau, Y.; Van de Peer, Y.; Rouzé, P.; Rombauts, S. PlantCARE, a database of plant cis-acting regulatory elements and a portal to tools for *in silico* analysis of promoter sequences. *Nucleic Acids Res.,* **2002**, *30*(1), 325-327. http://dx.doi.org/10.1093/nar/30.1.325 PMID: 11752327

[23] Mehrotra, R.; Mehrotra, S. Promoter activation by ACGT in response to salicylic and abscisic acids is differentially regulated by the spacing between two copies of the motif. *J. Plant Physiol.,* **2010**, *167*(14), 1214-1218. http://dx.doi.org/10.1016/j.jplph.2010.04.005 PMID: 20554077

[24] Sandelin, A.; Wasserman, W.W.; Lenhard, B. *ConSite: Web-based prediction of regulatory elements using cross-species comparison.,* **2004**.

[25] Basantani, M.; Gupta, D.; Mehrotra, R.; Mehrotra, S.; Vaish, S.; Singh, A. An update on bioinformatics resources for plant genomics research. *Curr. Plant Biol.,* **2017**, *2017*, 11. http://dx.doi.org/10.1016/j.cpb.2017.12.002

[26] Mehrotra, R.; Renganaath, K.; Kanodia, H.; Loake, G.J.; Mehrotra, S. Towards combinatorial transcriptional engineering. *Biotechnol. Adv.,* **2017**, *35*(3), 390-405. http://dx.doi.org/10.1016/j.biotechadv.2017.03.006 PMID: 28300614

[27] Sasaki, T.; Burr, B. International rice genome sequencing project: The effort to completely sequence the rice genome. *Curr. Opin. Plant Biol.,* **2000**, *3*(2), 138-141. http://dx.doi.org/10.1016/S1369-5266(99)00047-3 PMID: 10712951

[28] Rhee, S.Y.; Beavis, W.; Berardini, T.Z.; Chen, G.; Dixon, D.; Doyle, A.; Garcia-Hernandez, M.; Huala, E.; Lander, G.; Montoya, M.; Miller, N.; Mueller, L.A.; Mundodi, S.; Reiser, L.; Tacklind, J.; Weems, D.C.; Wu, Y.; Xu, I.; Yoo, D.; Yoon, J.; Zhang, P. The *arabidopsis* information resource (TAIR): A model organism database providing a centralized, curated gateway to *Arabidopsis* biology, research materials and community. *Nucleic Acids Res.,* **2003**, *31*(1), 224-228. http://dx.doi.org/10.1093/nar/gkg076 PMID: 12519987

[29] EMBL-EBI. The Gene Expression Atlas. Available from: http://www.ebi.ac.uk/gxa/.

[30] McDonald, J.H. *Handbook of Biological Statistics,* 2nd ed.; Baltimore, MD: sparky house publishing. **2009**.

[31] Bryne, J.C.; Valen, E.; Tang, M.H.E.; Marstrand, T.; Winther, O.; da Piedade, I.; Krogh, A.; Lenhard, B.; Sandelin, A. JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic Acids Res.,* **2008**, *36*(1), D102-D106. http://dx.doi.org/10.1093/nar/gkm955 PMID: 18006571

[32] Shao, H.; Wang, H.; Tang, X. NAC transcription factors in plant multiple abiotic stress responses: Progress and prospects. *Front. Plant Sci.,* **2015**, *6*, 902. http://dx.doi.org/10.3389/fpls.2015.00902 PMID: 26579152

[33] Huang, Q.; Wang, Y.; Li, B.; Chang, J.; Chen, M.; Li, K.; Yang, G.; He, G. TaNAC29, a NAC transcription factor from wheat, enhances salt and drought tolerance in transgenic *Arabidopsis. BMC Plant Biol.,* **2015**, *15*, 268. http://dx.doi.org/10.1186/s12870-015-0644-9 PMID: 26536863

[34] Krzywinski, M.; Altman, N. Visualizing samples with box plots. *Nat. Methods,* **2014**, *11*(2), 119-120. http://dx.doi.org/10.1038/nmeth.2813 PMID: 24645192

[35] Mehrotra, R.; Yadav, A.; Bhalothia, P.; Karan, R.; Mehrotra, S. Evidence for directed evolution of larger size motif in *Arabidopsis thaliana* genome. *Sci. World J.,* **2012**, *2012*, 983528. http://dx.doi.org/10.1100/2012/983528 PMID: 22645502