Comparative analysis of drought responsive calcium sensor genes and identification of global minima of calmodulin-1 by using conformation sampling approach

Lopus Merlin
Community Agro Biodiversity Center- M.S. Swaminathan Research Foundation, Wayanad- 673577, Kerala, INDIA
merlinlettizha@gmail.com

Abstract

Ca^{2+} is a crucial second messenger with response to wide variety of environmental stresses and development process. Out of three Ca^{2+} sensor gene families, two gene families (CBL and CaM) which act as sensor relays were selected for the study. The drought responsive members in CaM (Calmodulin) (LOC_Os03g20370) and CBL (Calcineurin B–like) (LOC_Os03g42840, LOC_Os05g45810, LOC_Os01g51420) were used for identifying their homologous genes in maize, sorghum and Arabidopsis. A total of 34 genes from the four species were analyzed for their phylogenetic relationship, gene structure, cis*-regulatory elements and expression level up on drought stress.*

Based on the comparative analysis, it was observed that all the identified homologous genes are potential for drought response which will be promising to the plant improvement programs. Further, rice calmodulin-1 was modeled by using the experimentally proven structure of homolog Arabidopsis calmodulin-7 and the native like structure was obtained through conformational sampling analysis.

Keywords: Drought, CBL, CaM, Conformational sampling, Homology model.

Introduction

Ca^{2+} is a crucial second messenger with response to wide variety of environmental stresses and development process[51]. Continuous sensing of changing and potentially harsh conditions induces various spatial and temporal patterns of Ca^{2+} levels in plants[8], these changes facilitate the plants to recognize the external stimuli that are decoded using highly-specific protein sensors which trigger the appropriate physiological responses. Ca^{2+} sensors affect the activity of downstream effectors that synchronize changes in metabolism, gene expression, and turnover of proteins[35].

In plants, there are three families of Ca^{2+} sensor proteins and they are: (i) calmodulin (CaM)/CaM-like (CML); (ii) calcineurin-B-like (CBL) and (iii) Ca^{2+}-dependent protein kinases (CDPKs, called CPKs in Arabidopsis)[7,13,65]. Out of three Ca^{2+} sensor families; CDPKs are the true "responders" that carry out direct signal transduction using their own catalytic activity. As CaMs/CMLs and CBLs lack catalytic activity they act as sensor relays which regulate downstream targets. Nevertheless, CBLs can specifically interact with CBL-interacting protein kinases (CIPKs) which are a specialized group of serine/threonine protein kinases[35].

CBL gene family has ten members in the genomes of *Arabidopsis* and rice. The ten CBL genes consists of six or seven introns in their coding regions of which four introns are conserved in the position and phase in all 10 CBL genes[30]. Each plant CBL has four EF-hands[3] and each EF-hand motif consists of 12 amino acid residues in which the amino acids at positions 1(X), 3(Y), 5(Z), 7(-X), 9(-Y) and 12(-Z) are highly conserved and responsible for binding of Ca^{2+}. The plant CBL harbors three canonical EF-hands where the first EF-hand motif is not canonical[3, 40]. Nagae et al analyzed the crystal structure of AtCBL2 protein and verified that the first and fourth EFhand motifs bind two calcium ions, respectively, whereas the second and third EF-hand motifs remain open.

CaM is the best characterized Ca^{2+} sensor and highly conserved from an evolutionary point and is present in all eukaryotic cells[4,19]. CaM is a small (149 amino acids) acidic protein and has a flexible helical region in the center which connects two globular domains. Each of these domains has two EF-hands that binds with Ca^{2+}. Even though multiple CaM isoforms are present, plant genomes encode a remarkable number of CMLs whose primary sequences have ≥16% overall identity with the canonical CaM sequence. Functional motifs other than EF-hands are notably absent[42,43].

In this study, drought responsive genes of CaM and CBL from rice have been compared with its homologous genes in maize, sorghum and *Arabidopsis* by using genomic analysis. They were analyzed for its common characteristics and also our study incorporates the possibility of conformational sampling in finding out the best structure for a homology model and various structural deviations of a protein in solvent.

Material and Methods

Data sets: The drought responsive members in CaM (Calmodulin) (LOC_Os03g20370) and CBL (Calcineurin B–like) (LOC_Os03g42840, LOC_Os05g45810, LOC_Os01g51420) gene families in rice were identified from previous studies[25,63]. Orthologs and paralogs of the selected genes were identified through BLASTp (https://blast.ncbi.nlm.nih.gov) search. Plant genome duplication database (PGDD)[36] was used for the

identification of segmental duplication. Sequences which had similarity >75%, were considered as potential orthologs by means of reciprocal best hit approach. Genomic, CDS and protein sequences of rice, maize, sorghum and Arabidopsis genes were retrieved from Rice Genome Annotation Proje*ct (*http://rice.plantbiology.msu.edu), Ensembl Plants (http://plants.ensembl.org) and NCBI (https://www.ncbi.nlm.nih.gov). Total of 34 genes were considered for the further analysis.

Multiple sequence alignment: Multiple sequence alignment was done for all the protein sequences by using ClustalW with MView output embedded in BAR (The Bio-Analytic Resource for Plant Biology) (https://bar.utoronto.ca/ntools/cgi-bin/ntools_multiplealign_w_mview.cgi).[61]

Phylogenetic analysis: The multiple sequence alignment of all the protein sequences was constructed by using CLUSTALW [57] and submitted to MEGA X [34] for building phylogenetic tree using Neighbour-Joining method by considering 1,000 rapid bootstrap replicates and it was visualized using iTOL (http://itol.embl.de). Discrete gamma distribution was analyzed by using the multiple sequence alignment to understand the evolutionary rate difference. Five discrete categories were used for the analysis. Substitution pattern and rates were estimated under the Jones-Taylor-Thornton model (+G)[24].

Gene structure analysis: The gene structures of all the genes were predicted by aligning the coding sequence with its corresponding genomic sequence by using GSDS 2.0 server (http:// gsds.cbi.pku.edu.cn). GSDS 2.0 is an improved version of GSDS and it supports two more widely used annotation formats providing more comprehensive support for annotation files.

Promoter sequence analysis: Cis-regulatory elements of all the sequences were identified from PLACE database [21]. This database identifies motifs in plant cis-acting regulatory DNA elements which were collected from previously published reports. In addition to the reported motifs in the original publication, their variations in other genes or other plant species reported later were also included.

Gene expression analysis: To understand the function of the genes, its expression pattern towards drought stress was analyzed by using GENEVESTIGATOR[22], which has a manually curated and well annotated database of expression data collected from variety of public repositories including Gene Expression Omnibus [6] and Array Express[49]. For rice, Affymetrix Rice Genome Array and mRNA- seq Gene Level Oryza Sativa (ref: MSU v7.0) platforms were used with water deficit microarray datasets OS-00143, OS-00195, OS-00230, OS-0369 (GSE57950, GSE78972, GSE92989, PRJNA306542). For maize, mRNA- Seq Gene Level Zea Mays (ref: AGPV4) platform was used with water deficit microarray datasets ZM-00049, ZM-00050, ZM-00076, ZM-00078, ZM00086 and ZM-0124 (GSE48507, GSE40070, GSE71723, E-MTAB-4297, E-MTAB-4325, E-MTAB-4198).

For *Arabidopsis*, mRNA- Seq Gene Level Arabidopsis thaliana platform was used with water deficit microarray dataset AT-00743 (E-MTAB-3279). Sorghum was analyzed manually by using water deficit microarray data set GSE80699.

Homology modeling and identification of global minima: The 3D structure of the protein sequence of the gene LOC_Os03g20370 was modeled by using Swiss Model workspace[28]. Conformational sampling approach was used to generate ensembles to expand the chances of identifying an energetic landscape that closely matched the input structures[26]. Normal Mode-based Simulation (NMSim) (http://www.nmsim.de) approach [31] performs three types of simulations viz. unbiased exploration of conformational space, pathway generation by a targeted simulation, and radius of gyration (RoG)-guided simulation. The RoG-guided simulation type was used here to generate the ensembles of protein structures of LOC_Os03g20370.

The parameters used for the rigid cluster decomposition were as follows: energy cutoff for hydrogen bonds (−1.0kcal/mol), method for placing hydrophobic constraints (3), cutoff for including hydrophobic constraints (0.35Å). The method chosen for the normal mode analysis was rigid cluster normal mode analysis and the distance cutoff for interactions between C-alpha atoms was set to 10Å. The parameters for the simulation were as follows: number of trajectories (1), number of simulation cycles (500), number of NMSim cycles (1), frequency of writing out conformations (1), side chain distortions (0.3), normal mode range (1–50), ROG mode (1), and step size (0.5Å).

Trajectories were visualized using VEGA ZZ package[50]. The ensemble of LOC_Os03g20370 protein structures was analyzed for their energetic contributions through Bayesian Analysis Conformation Hunt (BACH) algorithm (http://bachserver.pd.infn.it/in) in which the all-atom energy score was computed based on 1091 parameters. The BACH score was used to discriminate the global minima from the ensemble [11, 59].

Results and Discussion

Analysis of gene duplication and mapping of genes to chromosome locations: The gene OsCBL3 had 45% of orthologous genes in maize, 22% in sorghum and 33% in *Arabidopsis*. The gene OsCBL4 had 50% of orthologs in maize and 50% in sorghum. There were no orthologous genes for OsCBL10 in maize and *Arabidopsis* and its orthologs were identified only from sorghum (table 1). The gene OsCaM1-1 had 70% of its orthologs found in *Arabidopsis* 20% in maize and 10% in sorghum. The gene OsCBL3 was located in chromosome 3 and its paralogous gene OsCBL2 was found in chromosome 12.

Similarly, OsCBL4 located in chromosome 5 had its paralogous genes at chromosome 2 and they were OsCBL7 and OsCBL8. OsCaM1-1 was located in chromosome 3 and its paralogous genes were identified in chromosome 11 (OsCML2), chromosome 1 (OsCML1), chromosome 12 (OsCML3), chromosome 5 (OsCML9) and in chromosome 7 (OsCaM1-2). These genes and its paralogs might be evolved as a result of segmental duplication since they are located within the known genomic blocks[18] whereas, the gene OsCBL10 (chromosome 1) and OsCBL9 (chromosome 1) were paralogous genes and they might be evolved due to tandem duplication since they are located adjacent in the same chromosome[45]. The Ka/Ks ratio for the tandem duplicated genes were found to be 0.16 and for the segmental duplicated pairs it varied from 0 to 0.84.

Table 1
Rice genes and their homologous genes

Locus Id	Gene	Chromosome	Location	Paralogous genes	Orthologous genes
LOC_Os03g42840	OsCBL3	3	23888734..23895460	LOC_Os12g40510	Zm00001d033295 SORBI_3008G152800 Zm00001d030955 At4g26570 At5g55990 SORBI_3008G046500 Zm00001d023504 Zm00001d023506
LOC_Os05g45810	OsCBL 4	5	26535703..26537911	LOC_Os02g18880 LOC_Os02g18930	SORBI_3009G210300 Zm00001d038730
LOC_Os01g51420	OsCBl 10	1	29568428..29572133	LOC_Os01g39770	SORBI_3003G275000
LOC_Os03g20370	OsCaM1-1	3	9195526..9198964	LOC_Os01g59530 LOC_Os11g03980 LOC_Os12g03816 LOC_Os05g41200 LOC_Os07g48780	SORBI_3001G390300 At3g43810 Zm00001d038543 At2g27030 At3g56800 At2g41110 At5g21274 Zm00001d028948 At1g66410 At5g37780

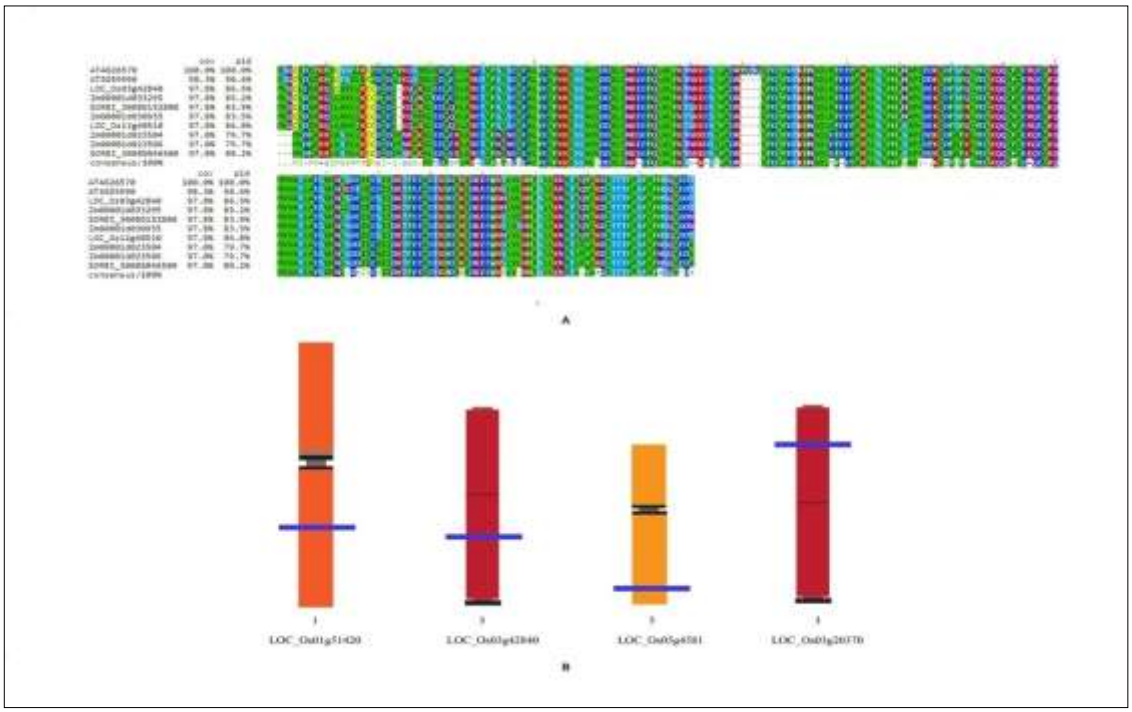

Fig. 1: A. Multiple sequence analysis of OsCBL3 and homologous genes, B. Chromosome location of the 4 selected genes

Analysis of domain architecture: The gene OsCBL3 (LOC_Os03g42840) and its homologous genes were analyzed for sequence conservation by multiple sequence alignment (figure 1). The protein sequence *Arabidopsis* gene At4g26570 was the reference and all the sequences showed percentage of identity ranging from 79.7 to 90.4%. The reference sequence AT4G26570 and AT5G55990 had 4 EF hand domains at the amino acid residues 36-81, 82-117, 119-154 and 163-197.

Similarly, the rice proteins sequence of the genes LOC_Os03g42840 and LOC_Os12g40510 had 4 EF hand domains at the amino acid residues 45-80, 81-116, 118-153, and 162-197 whereas in Zm00001d033295, SORBI_3008G152800 and Zm00001d030955 lacked EF hand 1 but the other three EF hands were highly conserved as in rice and *Arabidopsis*. Similarly, Zm00001d023504, Zm00001d023506 and SORBI_3008G046500 also lacked EF hand 1 domain and other three EF hand domains had slight change in amino acid residues when compared to remaining sequences. The position of three domains was at 79-114, 116-151 and 160-195. Except EF hand 1, other three domains had shown similarities in boundaries in all the sequences.

In the analysis of OsCBL4 (LOC_Os05g45810) and its homologous sequences, SORBI_3009G210300 was the reference sequence and the percentage of identity ranged from 64.1to 84.4% for the other sequences to the reference. The three rice sequences (LOC_Os05g45810, LOC_Os02g18880 and LOC_Os02g18930) had 4 EF hands at amino acid residues 31-66, 67-102 and 104-139. The other two sequences (SORBI_3009G210300 and Zm00001d038730) lacked the EF hand 1 but the three EF hands were highly conserved.

The OsCBL10 (LOC_Os01g51420) and its two homologous sequences were analyzed for conservation by taking LOC_Os01g51420 as reference. Percentage of identity ranged from 52.8 to 81.3%. Even though 4 EF hand domains were present in all the sequences, the domain boundaries varied slightly in all the three. OsCaM1-1 (LOC_Os03g20370) and its 15 homologous sequences were also analyzed by considering LOC_Os11g03980 as reference. The percentage of identity ranged from 38.7 to 99.33% for the members of the group to the reference.

The number of conserved EF hands was same in all the sequences but a slight change in boundaries was visible. Two rice sequences (LOC_Os11g03980 and LOC_Os12g03816) had 4 EF hands at amino acid residues 7-42, 43-78, 80-115 and 116-151 whereas other 4 rice sequences (LOC_Os07g48780, LOC_Os03g20370, LOC_Os01g59530 and LOC_Os05g41200), 5 *Arabidopsis* sequences (AT2G27030, AT3G56800, AT2G41110, AT5G21274, AT3G43810, AT1G66410 and AT5G37780) and 1 sorghum sequence SORBI_3001G390300 had EF hand domains at 8-43, 44-79, 81-116 and 117-149 amino acid residues.

One of the maize sequences in the group Zm00001d028948 had EF hand domains at 72-91, 92-127, 129-164 and 165-197 and the other sequence Zm00001d038543 had EF hand domains at 46-63, 64-99, 101-136 and 137-169 amino acid residues. The conservation in the domains and the similarities in the boundaries of homologous sequences showed that they might share similar structure and function[16].

Evolutionary analysis: The evolutionary relationship of the selected genes was analyzed after building a rooted phylogenetic tree (figure 2). Based on the tree, the sequences were grouped in to three. Group I (LOC_Os02g18880, LOC_Os02g18930, LOC_Os05g45810, SORBI_3009G210300 and Zm00001d038730) was OsCBL4 (LOC_Os05g45810) and its homologs. Group II (LOC_Os01g39770, LOC_Os01g51420, SORBI_3003G275000, SORBI_3008G046500, Zm00001d023504, Zm00001d023506, At4g26570, At5g55990, LOC_Os03g42840, Zm00001d033295, LOC_Os12g40510, SORBI_3008G152800 and Zm00001d030955) consisted of OsCBL10 (LOC_Os01g51420), OsCBL3 (LOC_Os03g42840) and their homologs. Group III (LOC_Os05g41200, LOC_Os01g59530, LOC_Os11g03980, LOC_Os12g03816, At5g21274, At3g56800, At2g41110, LOC_Os03g20370, LOC_Os07g48780, SORBI_3001G390300, At5g37780, At2g27030, At1g66410, Zm00001d028948, At3g43810, and Zm00001d038543) consisted of OsCaM1-1 (LOC_Os03g20370) and its homologs.

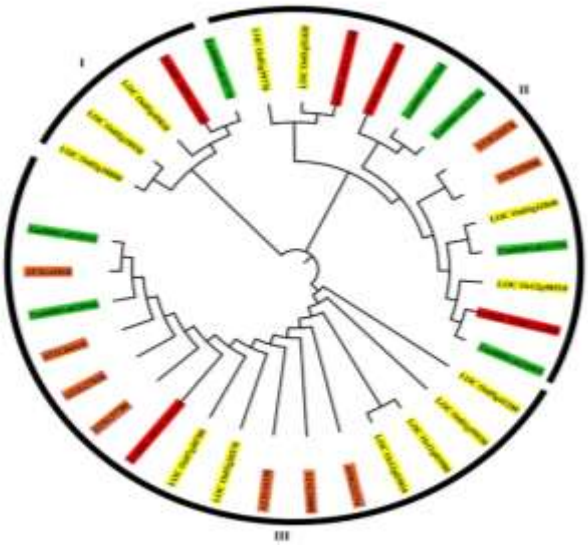

Fig. 2: Phylogenetic tree of 4 rice genes and homologous genes

Evolutionary rate was calculated by identifying the Gamma Parameter for site rates. Maximum likelihood was used as

the statistical method. The estimated value of the shape parameter for the discrete Gamma Distribution was 1.2651. A discrete Gamma distribution was used to model evolutionary rate differences among sites (5 categories, [+G]).

Mean evolutionary rates in these categories were 0.15, 0.43, 0.76, 1.23, 2.43 substitutions per site. The amino acid frequencies were 7.69% (A), 5.11% (R), 4.25% (N), 5.13% (D), 2.03% (C), 4.11% (Q), 6.18% (E), 7.47% (G), 2.30% (H), 5.26% (I), 9.11% (L), 5.95% (K), 2.34% (M), 4.05% (F), 5.05% (P), 6.82% (S), 5.85% (T), 1.43% (W), 3.23% (Y), and 6.64% (V). The maximum Log likelihood for this computation was -5246.772. The shape parameter $\alpha>1$ indicates a bell shaped distribution which points single-region mutation rates approximately normally distributed[32,33].

Genomic sequence analysis: The structures of exons and introns were well conserved in group I genes. All the genes consisted of 8 exons and 7 introns in this group (figure 3). In group II, OsCBL10 (LOC_Os01g51420) had 9 exons and 8 introns whereas its homologs LOC_Os01g39770 and SORBI_3003G275000 had 5 exons and 4 intons, 10 exons and 9 introns respectively. The variation in number of introns among the members of same group could have arisen due to selection pressure[41]. OsCBL3 (LOC_Os03g42840) and its 8 homologs had 8 exons 7 introns. The other homolog At4g26570 had 7 exons and 6 introns.

The intron richness of the group I and II genes indicates the functional diversity through alternate splicing and exon shuffling[27]. In group III, OsCaM1-1 (LOC_Os03g20370) and 10 of its homologs shared similar number of exons and introns (2 exons and 1 intron). Four homologs (LOC_Os11g03980, LOC_Os12g03816, At2g27030 and Zm00001d028948) had 3 exons and 2 introns. Two other homolgs At1g66410 and At3g43810 were with 1 exon and no introns.

In each phylogenetic group rice genes and majority of its homologs shared similarities in intron-exon structure while domains with conserved gene structure may involve in similar biological processes.[9] The homolgs were considered for functional analysis.

Drought related cis-regulatory elements in promoter sequences: The stress related cis-regulatory elements in the promoter sequences of the 4 rice genes and its homologs can provide information about environmental stimuli that might influence their gene expression. The identified drought related cis-regulatory elements (DRCR) in the sequences are given in table 2. All the rice genes and their homologs included drought related cis-regulatory elements (supplementary material I). The members in the group of OsCBL3 (LOC_Os03g42840) and homologs had 8 to 14 DRCR. The other group OsCBL 4 (LOC_Os05g45810) and homologs had DRCR in the range of 9 to 14. OsCBL 10 (LOC_Os01g51420) and its homologs had DRCR in the range of 10 to 13.

Similarly, the number of DRCR in OsCam1-1 (LOC_Os03g20370) and its homologs ranged from 5 to 13. Two maize genes, Zm00001d033295 and Zm00001d028948 had unique DRCRs and they were ABREATRD22 and ABREZMRAB28 respectively. Abscisic acid (ABA) in plants is accumulated under osmotic stress conditions caused by drought and ABA-responsive element (ABRE) is the major cis-regulatory element for ABA-responsive gene expression [47].

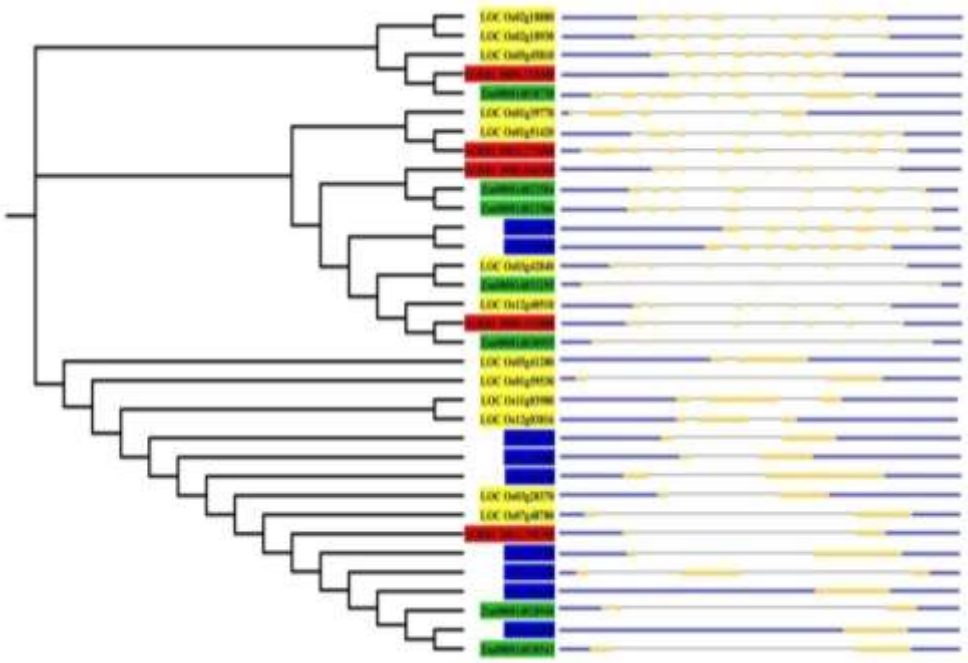

Fig. 3: Gene structure of 4 rice genes and homologous genes

Table 2
Drought related cis-regulatory elements identified in the genes

Drought related cis regulators	Accession No	PLACE ID
ACGT sequence[48]	S000415	ACGTATERD1
ABRE-like sequence[43]	S000414	ABRELATERD1
DRE2[49]	S000402	DRE2COREZMRAB17
DRE/CRT[50]	S000418	DRECRTCOREAT
CBF[51]	S000497	CBFHV
DRE1[52]	S000401	DRE1COREZMRAB17
MYB binding site[53]	S000408	MYB1AT
MYC binding site[54]	S000174	MYCATRD22
MYC binding site[55]	S000413	MYCATERD1
MYB binding site[56]	S000176	MYBCORE
MYB binding site[57]	S000177	MYB2AT
MYB binding site[54]	S000175	MYBATRD22
MYC binding site[58]	S000407	MYCCONSENSUSAT
MYB binding site[53]	S000409	MYB2CONSENSUSAT
ABRE[59]	S000133	ABREZMRAB28
ABRE[54]	S000013	ABREATRD22

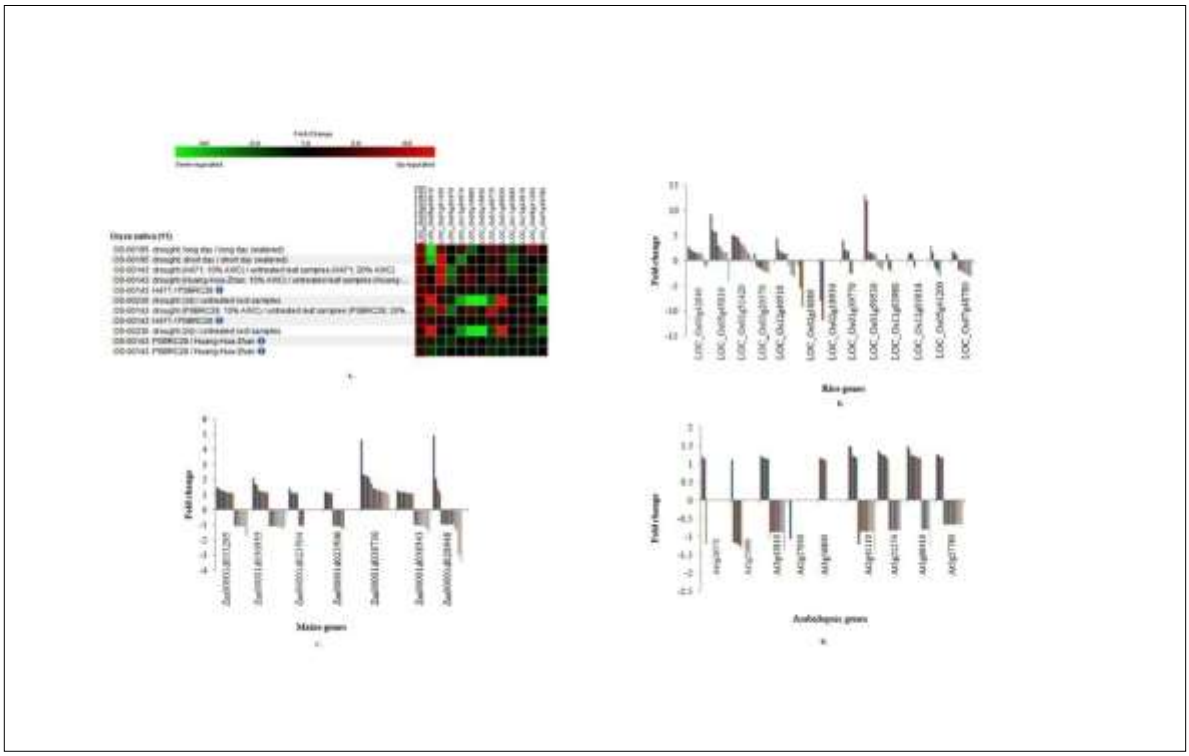

Fig. 4: A. Gene expression of LOC_Os03g42840 by using 11 data sets, B. Gene expression analysis of rice genes C. Gene expression analysis of maize genes, D. Gene expression analysis of *Arabidopsis* genes

Another identified promoter sequence DRE/CRT was included in genes showing ABA-independent expression in stress responses and tolerance. DREB1/CBF Transcription factors specifically interact with the DRE/CRT and control the expression of a large number of stress-responsive genes in *Arabidopsis*[38]. Rice DREB1/CBF-type transcription factors are involved in cold-responsive gene expression and also conferred improved tolerance to drought in transgenic rice[23]. The other stress relted cis-regulatory elements identified belonged to WRKY and basic leucine –zipper (bZIP) families.[1,2,5,10,12,14,17,29,52-54,56,58,60]

Gene expression with response to drought treatment: Transcripts of 13 rice genes were subjected to gene expression analysis based on the reported micro array studies (figure 4). The number of samples analyzed varied among genes, LOC_Os03g42840 (11 data sets), LOC_Os05g45810 (12 data sets), LOC_Os01g51420 (12 data sets), LOC_Os03g20370 (9 data sets), LOC_Os12g40510 (11 data

sets), LOC_Os02g18880 (3 data sets), LOC_Os02g18930 (2 data sets), LOC_Os01g39770 (6 data sets), LOC_Os01g59530 (12 data sets), LOC_Os11g03980 (3 data sets), LOC_Os12g03816 (4 data sets), LOC_Os05g41200 (6 data sets), LOC_Os07g48780 (12 data sets).

Except LOC_Os02g18880 and LOC_Os02g18930, all the other genes exhibited both up-regulation and down-regulation in response with various treatment of drought stress whereas LOC_Os02g18880 and LOC_Os02g18930 down-regulated in all the studies.

In the case of maize genes, samples used for each gene were as follows: Zm00001d033295 (16 data sets), Zm00001d030955 (17 data sets), Zm00001d023504 (8 data sets), Zm00001d023506 (10 data sets), Zm00001d038730 (15 data sets), Zm00001d038543 (17 data sets), Zm00001d028948 (15 data sets). The gene Zm00001d038730 showed up-regulation in all the studies whereas all the other genes showed up- regulation and down-regulation with response to various drought stress treatment. For *Arabidopsis*, samples used for each gene were as follows: At4g26570 (3 data sets), At5g55990 (6 data sets), At3g43810 (14 data sets), At2g27030 (1 data set), At3g56800 (4 data sets), At2g41110 (15 data sets), At5g21274 (13 data sets), At1g66410 (13 data sets) and At5g37780 (15 data sets).

The gene At3g56800 was up-regulated in all the 4 studies and the gene At2g27030, which was reported in only one study was found to be down-regulated. The other genes showed both up-regulation and down-regulation according to the drought stimulus. In sorghum 5 genes were analyzed manually by using the dataset GSE80699 and it was found that 4 genes (SORBI_3008G152800, SORBI_3008G046500, SORBI_3001G390300 and SORBI_3003G275000) were up-regulated for the drought treatment and the gene SORBI_3009G210300 showed down regulation for drought treatment.

Molecular modeling and conformational sampling analysis: Since certain nuclear domains, gene structures are similar even when protein sequence similarity is low, sequences can only be aligned with knowledge of protein tertiary structure [64]. Homology models demonstrate the topology of a given protein with deviations from experimental structures and which are normally with Cα coordinate root-mean square deviations (RMSD)[62]. Structure refinement methods focus to improve the accuracy of homology models toward experimental quality[15].

Conformational sampling is a common approach to search for structures that are closer to the true native state of the protein and identify those via suitable scoring functions.[37,39,48,55] Structure refinement is attained when the sampling generates conformations closer to the native state of the protein and when the scoring protocol can differentiate such conformations[55]. This protocol works best when selected ensemble subsets are averaged to match the nature of experimental structures and reduce scoring function noise[44,48].

Out of 4 rice proteins in this study, calmodulin-1 (LOC_Os03g20370) was selected for the structural analysis as one of its homolog gene At3g43810 (calmodulin-7) had experimentally identified structure with sequence identity 99.33%. From the template of calmodulin -7 (PDB ID: 4AQR) A, B chains were selected for the structure modeling of the rice protein calmodulin-1 (figure 5).

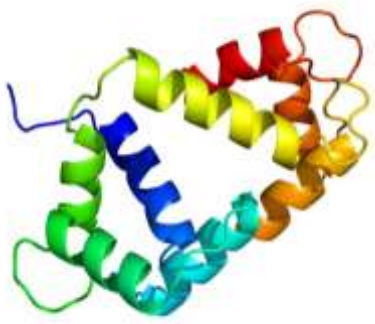

Fig. 5: Homology modeled structure of calmodulin-1

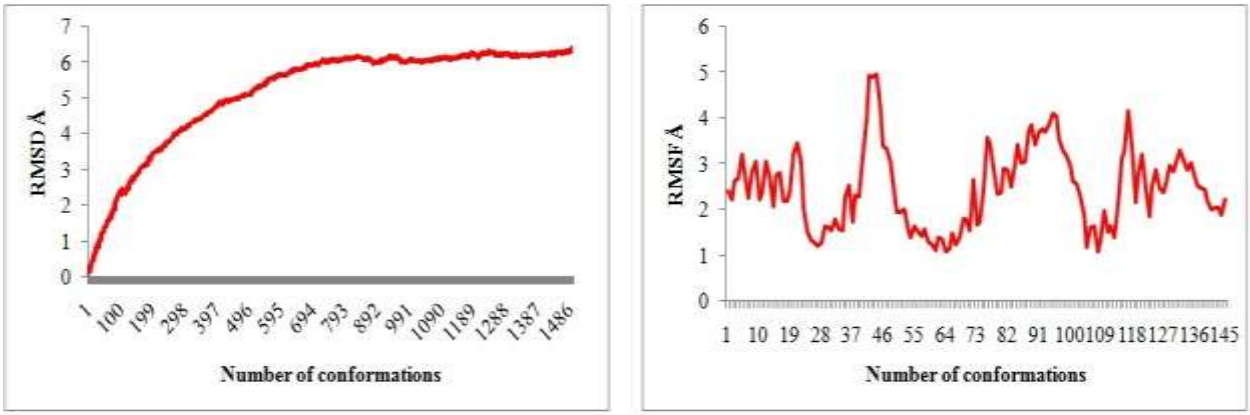

Fig. 6: RMSD and RMSF plots of calmodulin-1 by Conformational sampling analysis

The residue at most favored region according to ramachandran plot was 93.71%. QMEAN score for the modeled structure was 1.46 and MolProbity score was found to be 0.92 which indicates good agreement between modeled structure and experimental structure of similar size.

The structural diversity of the modeled structure was analyzed through NMSim program and the RMSD plot of Cα atoms was computed (figure 6) to understand the conformational space and rigidity of the ensemble and the average RMSD value was found to be 5.2 Å whereas overall flexibility of the residues or average RMSF value was 2.46 Å. Native like state of calmodulin-1 was identified by scrutinizing the global free energy minimum relative to all other states from the ensemble by using BACH algorithm [20].

Conclusion

Even though CBL and CaM family genes of calcium sensors are extensively studied for various stress responses, our study incorporates the possibility of conformational sampling in finding out the best structure for a homology model and various structural deviations of a protein in solvent. The native like structure was identified in rice calmodulin-1 by identifying the global minima from the ensemble. Common characteristics of 4 drought responsive rice genes and 30 of their homologous genes were identified in this study. The rice genes and their homologous genes shared sequence similarity ranging from 38.7 to 99.33%, even though slight boundary change was observed in OsCBL10, OsCaM1-1 and their homologs all the 4 EF hands were well conserved among them.

The orthologs from maize and sorghum lacked EF hand 1 in the case of OsCBL10 and OsCaM1-1. All the 4 rice genes and their ortholog *Arabidopsis* genes had well conserved 4 EF hand domains. The genes were further analyzed by their evolutionary relationship and it was found that OsCBL4 and its homologs clubbed to group I, OsCBL3 and OsCBL10 and their homologs in group II and OsCaM1-1 and its homologs were in group III. Similarly, in each phylogenetic group rice gene, majority of its homologs shared similarities in intron-exon structure.

Moreover, OsCBL3, OsCBL4 and OsCBL10 and their homologs exhibited intron richness which points to their functional diversity[27]. It was also observed that all the 34 genes had drought related cis-regulatory elements in their promoter sequence ranging from 5-14. Gene expression analysis proved that all the 34 genes are drought responsive as all the genes showed up-regulation or down-regulation upon various drought treatments. The results indicate that all the 30 homologous genes could be involved in drought response.

Acknowledgement

I would like to thank Kerala State Council for Science, Technology and Environment for the funding under Back to Lab programme (No751/2019/KSCSTE, 27/05/2019) and M. S. Swaminathan Research Foundation for providing necessary facilities to carry out this study. I thank Dr. Prajeesh Tomy, Assistant Professor VIT University, Vellore for his support to revise the manuscript.

References

1. Abe H., Urao T. and Ito T., Arabidopsis AtMYC2 (bHLH) and AtMYB2 (MYB) function as transcriptional activators in abscisic acid signaling, *Plant Cell*, **15**, 63-78 **(2003)**

2. Agarwal M., Hao Y. and Kapoor A., R2R3 type MYB transcription factor is involved in the cold regulation of CBF genes and in acquired freezing tolerance, *J Biol Chem*., **281**, 37636-37645 **(2006)**

3. Albrecht V., Ritz O. and Linder S., The NAF domain defines a novel protein-protein interaction module conserved in Ca2+-regulated kinases, *EMBO J*., **20**, 1051–63 **(2001)**

4. Astegno A. and Maresi E., Structural plasticity of calmodulin on the surface of caf2 nanoparticles preserves its biological function, *Nanoscale*, **6,** 15037–15047 **(2014)**

5. Banerjee A. and Roychoudhury A., Abscisic-acid-dependent basic leucine zipper (bZIP) transcription factors in plant abiotic stress, *Protoplasma*, **254**, 3–16 **(2017)**

6. Barrett T., Troup D.B. and Wilhite S.E., NCBI GEO: archive for functional genomics data sets–10 years on, *Nucleic Acids Res*., **39**, D1005-1010 **(2011)**

7. Batistic O. and Kudla J., Analysis of calcium signaling pathways in plants, *Biochim. Biophys. Acta*, **1820**, 1283–1293 **(2012)**

8. Bender K.W. and Snedden W.A., Calmodulin-related proteins step out from the shadow of their namesake, *Plant Physiol*., **163**, 486–495 **(2013)**

9. Betts M.J., Guigó R., Agarwal P. and Russell R.B., Exon structure conservation despite low sequence similarity: a relic of dramatic events in evolution?, *EMBO J*., **20**, 5354-60 **(2001)**

10. Busk P.K. and Pages M., Regulation of abscisic acid-induced transcription, *Plant Mol Biol.*, **37**, 425-435 **(1998)**

11. Cossio P., Granata D., Laio A., Seno F. and Trovato A., A simple and efficient statistical potential for scoring ensembles of protein structures, *Sci Rep*., **2**, 351 **(2012)**

12. Dubouzet J.G., Sakuma Y. and Ito Y., OsDREB genes in rice, Oryza sativa L., encode transcription activators that function in drought-, high-salt- and cold-responsive gene expression, *Plant J*., **33**, 751-763 **(2003)**

13. Edel K.H. and Kudla J., Increasing complexity and versatility: How the calcium signaling toolkit was shaped during plant land colonization, *Cell Calcium*, **57**, 231–246 **(2015)**

14. Fang Y., You J., Xie K., Xie W. and Xiong L., Systematic sequence analysis and identification of tissue-specific or stress-responsive genes of NAC transcription factor family in rice, *Mol Genet and Genomics*, **280**, 547–563 **(2008)**

15. Feig M., Computational structure refinement: Almost there, yet still so far to go, *Wiley Interdiscip Rev Comput Mol Sci*., **7**, e1307 **(2017)**

16. Gerlt J.A. and Babbitt P.C., Divergent evolution of enzymatic function:mechanistically diverse superfamilies and functionally distinct suprafamilies, *Annu. Rev. Biochem*., **70**, 209-246 **(2001)**

17. Gilmour S.J., Zarka D.G. and Stockinger E.J., Low temperature regulation of the Arabidopsis CBF family of AP2 transcriptional activators as an early step in cold-induced COR gene expression, *Plant J*., **16**, 433-442 **(1998)**

18. Guo H., Lee T.H. and Wang X., Function relaxation followed by diversifying selection after whole-genome duplication in flowering plants, *Plant Physiol*., **162**, 769–778 **(2013)**

19. Halling D.B., Liebeskind B.J., Hall A.W. and Aldrich R.W., Conserved properties of individual ca2+-binding sites in calmodulin, Proc. Natl. Acad. Sci., USA, **113**, E1216–E1225 **(2016)**

20. Heo L. and Feig M., Experimental accuracy in protein structure refinement via molecular dynamics simulations, Proc Natl Acad Sci., USA, **115**, 13276-13281 **(2018)**

21. Higo K., Ugawa Y., Iwamoto M. and Korenaga T., Plant cis-acting regulatory DNA elements (PLACE) database, *Nucleic Acids Research*, **27**, 297-300 **(1999)**

22. Hruz T., Laule O., Szabo G. and Wessendorp F., Genevestigator v3: a reference expression database for the meta-analysis of transcriptomes, *Adv Bioinformatics*, **10**, 420747-420757 **(2008)**

23. Ito Y., Katsura K., Maruyama K. and Taji T., Functional analysis of rice DREB1/CBF-type transcription factors involved in cold-responsive gene expression in transgenic rice, *Plant Cell Physiol*., **47**, 141–153 **(2006)**

24. Jones D.T., Taylor W.R. and Thornton J.M., The rapid generation of mutation data matrices from protein sequences, *Computer Applications in the Biosciences*, **8**, 275-282 **(1992)**

25. Kanwar P., Sanyal S.K., Tokas I., Yadav A.K. and Pandey A., Comprehensive structural, interaction and expression analysis of CBL and CIPK complement during abiotic stresses and development in rice, *Cell Calcium*., **56**, 81-95 **(2014)**

26. Kellogg E.H., Leaver-Fay A. and Baker D., Role of conformational sampling in computing mutation-induced changes in protein structure and stability, *Proteins Struct Funct Bioinformatics*, **79**, 830–840 **(2011)**

27. Keren H., Lev-Maor G. and Ast G., Alternative splicing and evolution: diversification, exon definition and function, *Nat. Rev. Genet*., **11**, 345–355 **(2010)**

28. Kiefer F., Arnold K., Künzli M., Bordoli L. and Schwede T., The SWISS-MODEL repository and associated resources, *Nucleic Acids Res*., **37**, D387–D392 **(2009)**

29. Kizis D. and Pages M., Maize DRE-binding proteins DBF1 and DBF2 are involved in rab17 regulation through the drought-responsive element in an ABA-dependent pathway, *Plant J*., **30**, 679-689 **(2002)**

30. Kolukisaoglu U., Weinl S. and Blazevic D., Calcium sensors and their interacting protein kinases: genomics of the Arabidopsis and rice CBL–CIPK signaling networks, *Plant Physiol*., **134**, 43–58 **(2004)**

31. Kruger D.M., Ahmed A. and Gohlke H., NMSim web server: integrated approach for normal mode-based geometric simulations of biologically relevant conformational transitions in proteins, *Nucleic Acids Res*., **40**, W310–6 **(2012)**

32. Kuhner M.K., LAMARC 2.0: maximum likelihood and Bayesian estimation of population parameters, *Bioinformatics*, **22**, 768-770 **(2006)**

33. Kuhner M.K. and Smith L.P., Comparing Likelihood and Bayesian Coalescent Estimation of Population Parameters, *Genetics*, **175**, 155-165 **(2007)**

34. Kumar S., Stecher G., Li M., Knyaz C. and Tamura K., MEGA X: Molecular Evolutionary Genetics Analysis across computing platforms, *Molecular Biology and Evolution*, **35**, 1547-1549 **(2018)**

35. La Verde V., Dominici P. and Astegno A., Towards Understanding Plant Calcium Signaling through Calmodulin-Like Proteins: A Biochemical and Structural Perspective, *Int J Mol Sci*., **19**, 5 **(2018)**

36. Lee T.H., Kim J., Robertson J.S. and Paterson A.H., Plant Genome Duplication Database, *Methods Mol Biol*., **1533**, 267-277 **(2017)**

37. Lin M.S. and Head-Gordon T., Reliable protein structure refinement using a physical energy function, *J Comput Chem*., **32**, 709–717 **(2011)**

38. Liu Q., Kasuga M. and Sakuma Y., Two transcription factors, DREB1 and DREB2, with an EREBP/AP2 DNA binding domain separate two cellular signal transduction pathways in drought- and low-temperature-responsive gene expression, respectively, in Arabidopsis, *Plant Cell*, **10**, 1391–1406 **(1998)**

39. Lu H. and Skolnick J., Application of statistical potentials to protein structure refinement from low resolution ab initio models, *Biopolymers*, **70**, 575–584 **(2003)**

40. Luan S., Kudla J. and Rodriguez-Concepcion M., Calmodulins and calcineurin B-like proteins: calcium sensors for specific signal response coupling in plants, *Plant Cell*, **14**, 389–400 **(2002)**

41. Lynch M., Intron evolution as a population-genetic process, Proc. Natl. Acad. Sci., USA, **99**, 6118–6123 **(2002)**

42. McCormack E. and Braam J., Calmodulins and related potential calcium sensors of Arabidopsis, *New Phytol*., **159**, 585–598 **(2003)**

43. McCormack E., Tsai Y.C. and Braam J., Handling calcium signaling: Arabidopsis cams and cmls, *Trends Plant Sci*., **10**, 383–389 **(2005)**

44. Mirjalili V. and Feig M., Protein structure refinement through structure selection and averaging from molecular dynamics ensembles, *J Chem Theory Comput.*, **9**, 1294–1303 **(2013)**

45. Mittal S. et al, Comparative Analysis of CDPK Family in Maize, Arabidopsis, Rice, and Sorghum Revealed Potential Targets for Drought Tolerance Improvement, *Front. Chem.*, **5**, 115 **(2017)**

46. Nagae M., Nozawa A. and Koizumi N., The crystal structure of the novel calcium-binding protein AtCBL2 from Arabidopsis thaliana, *J Biol Chem*., **278**, 42240–6 **(2003)**

47. Nakashima K., Fujita Y. and Katsura K., Transcriptional regulation of ABI3- and ABA-responsive genes including RD29B and RD29A in seeds, germinating embryos, and seedlings of Arabidopsis, *Plant Mol Biol.*, **60**, 51-68 **(2006)**

48. Park H., DiMaio F. and Baker D., The origin of consistent protein structure refinement from structural averaging, *Structure*, **23**, 1123–1128 **(2015)**

49. Parkinson H., Sarkans U. and Kolesnikov N., Array Express update–an archive of microarray and high-throughput sequencing-based functional genomics experiments, *Nucleic Acids Res.*, **39**, D1002-1004 **(2011)**

50. Pedretti A., Villa L. and Vistoli G., VEGA–an open platform to develop chemo-bio-informatics applications, using plug-in architecture and script programming, *J Comput Aided Mol Des*., **18**, 167–73 **(2004)**

51. Sanders D., Brownlee C. and Harper J.F., Communicating with calcium, *Plant Cell*, **11**, 691–706 **(1999)**

52. Simpson S.D., Nakashima K. and Narusaka Y., 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.*, **33**, 259-270 **(2003)**

53. Skinner J.S., von Zitzewit J. and Szucs P., Structural, functional, and phylogenetic characterization of a large CBF gene family in barley, *Plant Mol Biol.*, **59**, 533-551 **(2005)**

54. Solano R., Nieto C. and Avila J., Dual DNA binding specificity of a petal epidermis-specific MYB transcription factor (MYB.Ph3) from Petunia hybrid, *EMBO J*., **14**, 1773-1784 **(1995)**

55. Stumpff-Kane A.W., Maksimiak K., Lee M.S. and Feig M., Sampling of near-native protein conformations during protein structure refinement using a coarse-grained model, normal modes, and molecular dynamics simulations, *Proteins*, **70**, 1345–1356, **(2008)**

56. Svensson J.T., Crosatti C. and Campoli C., Transcriptome analysis of cold acclimation in barley albina and xantha mutants, *Plant Physiol.*, **141**, 257-270 **(2006)**

57. Thompson J.D., Higgins D.G. and Gibson T.J., CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, *Nucleic Acids Res*., **22**, 4673–4680 **(1994)**

58. Tran L.S., Nakashima K. and Sakuma Y., 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*, **16**, 2481-2498 **(2004)**

59. Trosset J.Y. and Scheraga H.A., Reaching the global minimum in docking simulations: a Monte Carlo energy minimization approach using Bezier splines, Proc Natl Acad Sci., USA, **95**, 8011–5 **(1998)**

60. Urao T., Yamaguchi-Shinozaki K., Urao S. and Shinozaki K., An Arabidopsis myb homolog is induced by dehydration stress and its gene product binds to the conserved MYB recognition sequence, *Plant Cell*, **5**, 1529-1539 **(1993)**

61. Waese J. and Provart N.J., The Bio-Analytic Resource for Plant Biology, *Methods Mol Biol.*, **1533**, 119-148 **(2017)**

62. Waterhouse A., SWISS-MODEL: Homology modelling of protein structures and complexes, *Nucleic Acids Res.*, **46**, W296–W303 **(2018)**

63. Wu H.C. and Jinn T.L., Oscillation regulation of Ca2+ /calmodulin and heat-stress related genes in response to heat stress in rice (Oryza sativa L.), *Plant Signal Behav*., **7**, 1056-7 **(2012)**

64. Zhang J., Liang Y. and Zhang Y., Atomic-level protein structure refinement using fragment-guided molecular dynamics conformation sampling, *Structure*, **19**, 1784–1795 **(2011)**

65. Zhu X., Dunand C., Snedden W. and Galaud J.P., Camand cml emergence in the green lineage, *Trends Plant Sci.*, **20**, 483–489 **(2015)**.

(Received, accepted)