Australian Journal of Crop Science AJCS

AJCS 10(2):207-215 (2016) ISSN:1835-2707

Cloning and comparative protein modelling of two MADS-box genes, HsMADS1 and HsMADS2 isolated from Hibiscus sabdariffa L. var. UMKL (roselle)

Siti N. Othman[1], Yong S.Y.C[1*], Roghayeh Abedi Karjiban[2], Adibah Shakri[3]

[1]Department of Biology, Faculty of Science, Universiti Putra Malaysia (UPM), 43400 Serdang, Selangor, Malaysia
[2]Department of Chemistry, Faculty of Science, Universiti Putra Malaysia (UPM), 43400 Serdang, Selangor, Malaysia
[3]Institute of Bioscience, Universiti Putra Malaysia (UPM), 43400, Serdang, Selangor, Malaysia

***Corresponding author: chrisyong@upm.edu.my**

Abstract

Hibiscus sabdariffa L. var. UMKL or commonly known as roselle is cultivated in Malaysia mainly for its calyx, which is high in vitamin C and anthocyanin. Unfortunately, the genetic information regarding the flowering pathway of roselle is very scarce. It is essential to understand the genetics underlying roselle's flower developmental process by studying MADS-box transcription factor genes that play crucial roles in controlling the development of calyx in flowering plants. Designated as *HsMADS1* and *HsMADS2*, two MADS-box genes were isolated from the calyx tissues of roselle from different developmental stages using 3'- RACE PCR and primer walking approaches. The different motifs in the C domain region of *HsMADS1* and *HsMADS2* deduced amino acid sequences suggested that both genes probably originated from *SEP* and *AGL6* subfamilies of MADS-box gene respectively. The putative functions of the genes based on BLAST searches and phylogenetic analyses suggested that *HsMADS1* possibly involves in the expression of *SEP* gene in stem, leaf, bud and flower organs of roselle, whereas *HsMADS2* may probably involve in the late expression of floral tissue for stem branching. The alpha helix rich structures of SRF-TF identified in the deduced amino acid sequences of HsMADS1 and HsMADS2 supported the involvement of both proteins in DNA binding and dimerisation.

Keywords: Gene isolation, *in silico* analyses, molecular modeling, flower development, Transcriptional Factor.
Abbreviations: *AGL6*_ Agamous-like 6; *SEP*_ SEPALLATA; SRF-TF_ Serum Response Factor – Trancriptional Factor.

Introduction

Hibiscus sabdariffa L. var. UMKL is widely utilised in food, beverages and pharmaceutical industries as synonym with its high content of vitamin C and anthocyanin (Mohamad et al. 2009). The calyx of its flower is incontrovertibly the most profitable and considerable part in roselle plant. MADS-box genes had been reported to play critical roles in plant development and flower formation. Extensive studies on the characterization and expression of MADS-box genes involved in flower developmental of other economically important plant species such as *Coffea arabica*, L. (de Oliveira et al. 2014), *Crocus sativus* (Tsaftaris et al. 2005), wheat (*Triticum aestivum* L.) (Zhao et al. 2006) and *Alpinia hainanensis* (Song et al. 2010) have been conducted. However, there are no reported study regarding the MADS-box genes in roselle despite the commercial potentials of this plant. Many studies have been conducted on roselle but they were mainly focussing on physico-chemical properties and antioxidant content of the calyx (Da-Costa-Rocha et al. 2014; Mgaya-Kilima et al. 2015). Apart from the study on mutation breeding (Osman et al. 2011), there is a lack of molecular study on the flowering genes in roselle. Very limited knowledge is available regarding the MADS-box genes that may be involved in its flowering pathway. Furthermore, most MADS-box genes isolated thus far from the different tissues of various species of angiosperms are present in isoforms. It would be interesting to isolate these isoforms and investigate whether or not the MADS-box genes isolated from the same type of tissue at different developmental stages present in isoforms in roselle. Since MADS-box genes are crucial in regulating the flowering pathway in angiosperm and calyx of roselle is highly sorted after for its pharmacological and nutritional values, two MADS-box genes were isolated from the calyx tissues of roselle at different developmental stages in this study. Fundamental knowledge about the primary gene structures, domains functional regions, protein structural motifs and the protein structures of the MADS-box genes are crucially needed for the understanding of the flowering pathway in *H. sabdariffa*. The isolation of the two MADS-box genes from the calyx tissue of roselle in this study provides fundamental genetic information that are useful for further study aiming at improving crop yields through genetic modifications. Besides that, this study also enhances the impact of structure and function on plant biology as it provided the predicted three dimensional (3D) structure of plant type MADS-box transcriptional factor and the possible molecular functions of the protein in plant based on the 3D structure.

Results

Cloning of CDS of HsMADS1 and HsMADS2

HsMADS1 (Accession number: KP942778) and *HsMADS2* (Accession number: KP942779) CDS were consisted of 951

bp and 981 bp, respectively. The start and stop codons, 3' UTR, poly-A signal and poly-A tail were identified in both sequences and both HsMADS1 and HsMADS2 encoded for putative proteins of 244 amino acids. Using BLAST P, the deduced amino acid sequence of *HsMADS1* was found to be 84% significantly identical with MADS-17 of *Gossypium hirsutum* (accession no: AEJ76841.1) whereas *HsMADS2* was 95% homologous with MADS-13 protein of *G. hirsutum* (accession no: FJ409870.1). In addition, the poly-A signal of *HsMADS1* and *HsMADS2* CDS were predicted at base pair 788 to 795 and base pair 852-862, respectively. Deduced amino acid sequences comparison between *HsMADS1* and *HsMADS2* CDS demonstrated only 44.5% of similarities. As a matter of interest, despite the dissimilarities in the nucleotide and deduced amino acid sequences of *HsMADS1* and *HsMADS2*, both genes encoded for ORF of 735 bp. The isolated *HsMADS1* and *HsMADS2* contained all the domains of a typical plant MADS box gene, which consisted of an MADS domain, K domain, a short I region, and the C terminal region. MADS and K domains were predicted by Interproscan and Pfam software. These two distinctive domains predicted in the HsMADS1 and HsMADS2 proteins have a specific transcription factor namely the Serum Response Factor-Transcriptional Factor (SRF-TF) as the signatures of their domains. I and C domains were predicted through multiple sequences comparison of these two genes with their respective MADS-box homologous sequences from other species. The deduced amino acid sequences of *HsMADS1* and *HsMADS2* were less conserved in the C region. Interestingly, based on comparative analysis with other homologous MADS-box sequences, two different conserved motifs were identified in the C region of the deduced amino acid sequences of *HsMADS1* and *HsMADS2*, respectively. Based on the multiple sequence alignment of *HsMADS1* with other MADS-box genes isolated from other plant species, two conserved amino acids CNPTLQIGY (internal) and GFIPGWML (terminal) motifs were predicted at the C terminal of HsMADS1 protein sequence. Meanwhile, two conserved CDHEPVLQIGY (internal) and FIHGWVI (terminal) motifs were also identified from HsMADS2 protein sequence (Fig. 1).

Isolation of the intronic regions

The assembly of the CDS and intronic sequences for both *HsMADS1* and *HsMADS2* generated gene sequences of 3,050 bp and 2,791 bp, respectively. Exon-Intron junctions of *HsMADS1* and *HsMADS2* genes were identified through comparison with their respective CDS sequences and supported by the comparison with homologous MADS-box genes from other species. The assembled nucleotides sequences of *HsMADS1* and *HsMADS2* were also aligned for a comparison and the percentage of nucleotides similarity was 42.46%. *HsMADS1* has seven exons while *HsMADS1* has eight exons (Fig. 2). Similar to *HsMADS1*, the four domains which were MADS, K, I and C domains were predicted in the deduced amino acids of *HsMADS2*. Through multiple sequence alignment of *HsMADS2* and *HsMADS1* with their homologous sequences, the MADS domain in respective *HsMADS1* and *HsMADS2* was predicted to be entirely encoded by the first exon. Meanwhile, the I domain of the two genes are entirely encoded by their second exons. Whereas, the K domain identified in *HsMADS2* (third to sixth exons) is encoded by more exons than *HsMADS1* (third to fifth exons), and the C domain was found in the last two exons of both genes.

Fig 1. Amino acids sequences comparison of HsMADS1 and HsMADS2 and related MADS-box proteins. A) Multiple alignment of deduced amino acids of HsMADS1 and its closest homologous. CNPTLQIGY (internal) and GFIPGWML (terminal) SEP motifs were boxed with red solid box labelled with I and II, respectively. The accession numbers are: GhMADS17(AEJ76841.1); GhMADS62 (AGW23364.1); TcMADS (XP_007032865.1); CpMADS1 (ACD39982.1) and CsSEP1 (XP_006482430.1). B) Multiple alignment of deduced amino acids of HsMADS2 and its closest homologous. CDHEPVLQIGY (internal) and FIHGWVI (terminal) AGL6 motifs are boxed by solid, yellow box labelled with I and II, respectively. The accession numbers are: GhMADS13 (ACJ26768.1), GhMADS45 (AGW23347.1) and GhMADS47 (AGW23348.1), HcMADS -box (ADZ98838.1) and TcAGL6 (XP_007051983.1). The conserved MADS-box domain is marked with solid line while conserved K-box domain is boxed with dashed line box. The residues in MADS region that have been coloured by yellow indicated the initiation and termination residues of the SRF-TF in the MADS domains, while the residues in K-box domain that have been coloured with red indicated the initiation and termination of K-domain transcriptional factor. The similar deduced amino acid sequences are shaded in black. The same nucleotides were indicated by the asterisks (*). The highly similar nucleotides indicated by colon (:) symbol. The more or less similar nucleotides indicated by the dot (.) symbol.

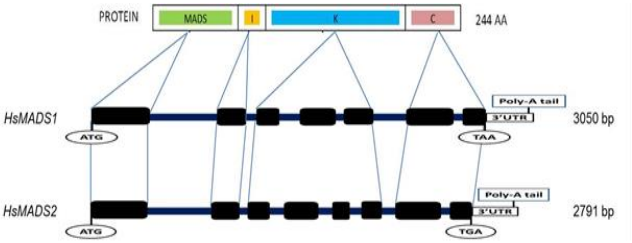

Fig 2. The structures of HsMADS1 and HsMADS2 genes. HsMADS1 has seven exons and six introns whereas HsMADS2 consisted of eight exons and seven introns. The coding region started at the start codon of respective genes. The boxes represent exons with open reading frame in black and the solid lines represent introns. The four characteristic domains are matched with the associated exons in HsMADS1 and HsMADS2. HsMADS1: First exon (190 bp), second exon (77 bp), third exon (70 bp), fourth exon (100 bp), fifth exon (80 bp), sixth exon (139 bp) and seventh exon (84 bp). HsMADS2: First exon (180 bp), second exon (57 bp), third exon (51 bp), fourth exon (98 bp), fifth exon (42 bp), sixth exon (41 bp), seventh exon (153 bp) and eighth exon (83 bp).

Phylogenetic tree

Neighbour joining tree was constructed with bootstrapped (n=2000 replicates) with 37 homologous MADS sequences and *AtAP1* as the out-group (Fig. 3). The clustering in the phylogenetic analysis showed that *HsMADS1* and *HsMADS2* are closely related to *SEP* and *AGL6* subfamilies of MADS-box genes. Evolutionary relationship showed that *HsMADS1* is closely related to *SEP* genes of *Gossypium arboretum*, *MADS17* of *G. hirsutum* and *MADS62* of *G. hirsutum* with high bootstrapped value. Meanwhile, *HsMADS2* was clustered together with *AGL6* genes from *Arabidopsis thaliana*, *SEP1* of *Agapanthus praecox* and *MADS6* of *Nelumbo nucifera* supported by strong bootstrapped value. The phylogenetic analysis on the MADS-box genes revealed that orthologous MADS-box genes are more similar compared to paralogous MADS-box genes in roselle.

Comparative Modelling of* HsMADS1 *and* HsMADS2 *using MODELLER ver 9.13

The accuracy of the predicted structures constructed by MODELLER ver 9.13 ranged from good to poor as sequence similarity of the predicted models with the templates (4OX0; 1EGW; 1N6J) with highest similarity percentages of 60% to lowest of 38%. Model 1 of HsMADS1 (HsMADS1. B99990001) protein structure was selected as the most favourable model due to its lowest score of DOPE which was -12034.94531 and with acceptable score of GA341, which was 0.7 (Fig. 4A). With the sequence length of 244 amino acid residues, the sequence identity of the target model and template was 66.7%. The secondary structures HsMADS1 predicted to consist of 48.0% of alpha helix, 31% disordered with 4.0% of beta-strands (Supplementary Fig. 1). The analysis of the stereochemical errors for HsMADS1 model confirmed that HsMADS1.B99990001 was a reasonable model. Ramachandran plot analyses also showed that 234 out of 244 (94.7%) residues of HsMADS1 were located in the favourites region. Besides that, 4 out of 244 (1.7%) residues located in

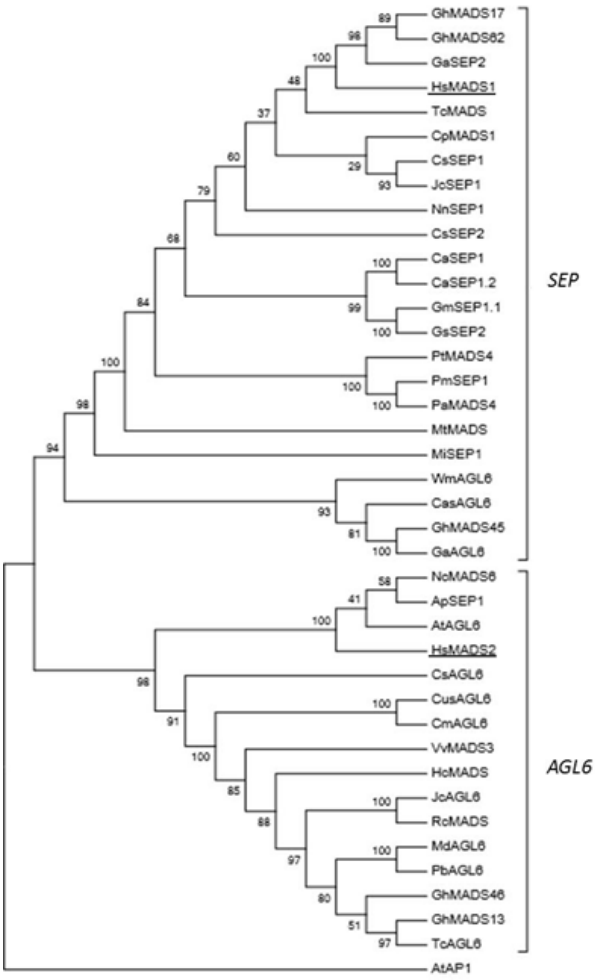

Fig 3. Neighbour-Joining analysis of representative plant E-function MADS-box genes with 36 homologous sequences from various angiosperms. The bootstrap values are expressed as a percentage (over 2000 replicates) and being shown at the corresponding nodes. The tree was clustered to *SEP* and *AGL6* groups.

the allowed region, and just 4 out of 244 (1.7%) located in the outlier region (Supplementary Fig. 2). Further stereochemical analyses by ProSA showed that overall Z-score of HsMADS1 was -0.28. The Z-score plot was located at the ideal range of Z-score from native protein of similar size with HsMADS1 from protein database (Supplementary Fig. 3). With the sequence length of 244 amino acids, the structure second model of HsMADS2 (HsMADS2. B99990002) was selected as the most significant model structure with the lowest of DOPE score -9234.760742 (below -1.0) and the highest GA341 score of 0.4 (near by 1.0) (Fig. 4B). The percentage of sequence identity for this model with its 1EGW template was 54.9%. The secondary structure of HsMADS2 was predicted to consist of 48.0% of alpha helix with 31.0% of disordered and 4.0% of beta-strands (Supplementary Fig. 4). Ramachandran plot result proved that out of the 244 residues of HsMADS2, 93% of the residues were located in favoured region (225 residues). Percentage of residues in allowed region was 4.1% (10 residues) and the percentage of residues in outlier region was 2.9% (7 residues) (Supplementary Fig. 5). The analysis of the stereochemical errors for HsMADS2 model with the ProSA showed the overall Z-score of HsMADS2.B99990002

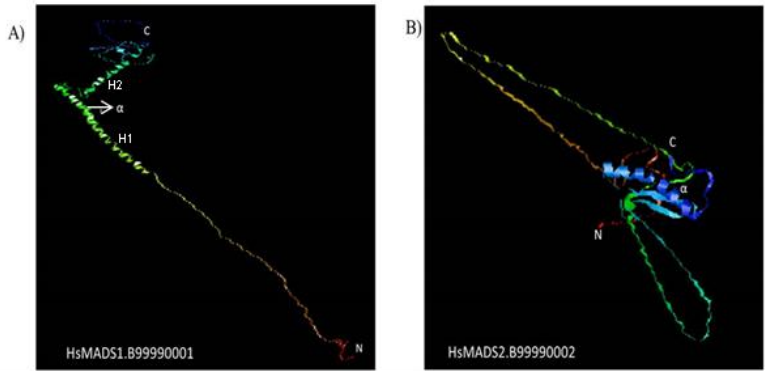

Fig 4. Comparative modelling structures for HsMADS1 and HsMADS2 proteins via Modeller ver 9.13. A) The comparative model of HsMADS1.B99990001. Protein structure of HsMADS1 with N-terminal, two alpha helixes (H1 and H2) and C-terminal. B) The comparative model of HsMADS2.B99990002. Protein structure of HsMADS2 with N-terminal, one alpha helix (α) and C-terminal (C). The ribbon structure was visualised using RASMOL.

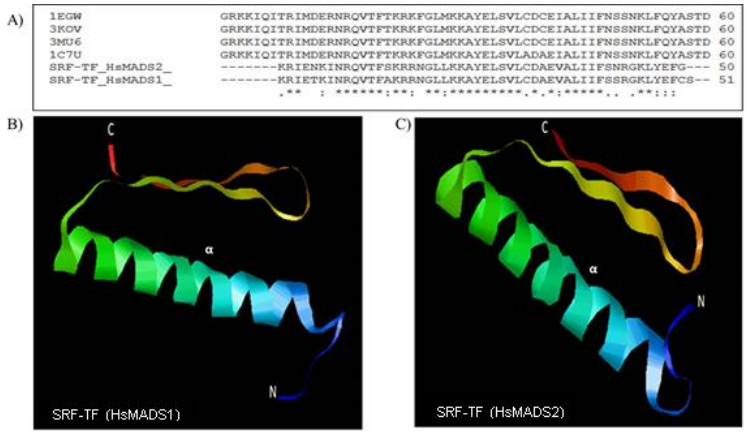

Fig 5. Comparative modelling structures of SRF-TF in MADS domains of HsMADS1 and HsMADS2 with the secondary structures prediction. A) Multiple sequences alignment of SRF-TF domain from HsMADS1 (9-59 aa) and SRF-TF domain from HsMADS2 (9-58 aa) with the templates using Clustal W2. B) The ribbon structure of the SRF-TF domain of HsMADS1 protein. SRF monomers indicated in blue and red in colours while with N-terminal, alpha helix (H1) and C-terminal were labelled. C) Comparative modelling of the specific SRF-TF at MADS domain of HsMADS2. The ribbon structure of the SRF-TF domain of *HsMADS2* protein. SRF monomers indicated in blue and red in colours while with N-terminal, alpha helix (H1) and C-terminal were labelled. The models are visualised using RASMOL. The elements presented in the secondary structure (c = random coil, h=Alpha helix, e=extended strands).

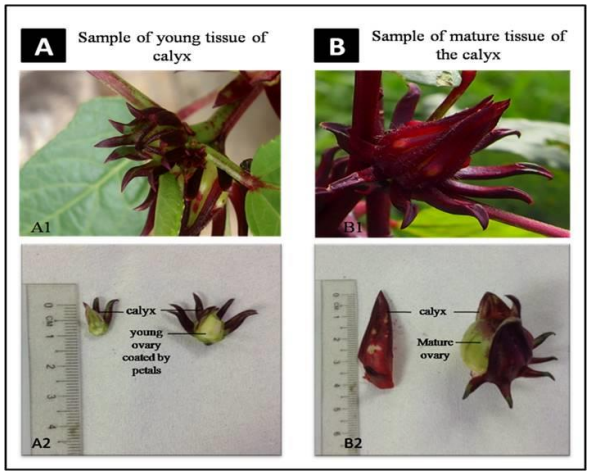

Fig 6. The young and mature calyces that were sampled for DNA and RNA extraction purposes. A1-A2; young calyx was sampled 8 days before anthesis (day 53 of the planting). B1-B2; mature calyx sampled a day post anthesis (day 61 of the planting).

model was 1.29. In Overal Model Quality plot, the Z-score of HsMADS2.B99990002 model was not in the range of the Z-score that typically found from native protein of similar size. The local quality graph showed that most of the residues of the model were in positive values and became the problematic parts of the input structure (Supplementary Fig. 6).

Serum Response Factor – Transcriptional Factor (SRF-TF) is one of the signature of the MADS-box domain identified in HsMADS1 and HsMADS2 proteins. The topology structure of SRF-TF domain for HsMADS1 is composed of one alpha (α) helix. The secondary structure of SRF-TF of HsMADS1 as predicted by Garnier-Osgurthorpe-Robson (GOR) server were consisted of 62.75 % of α helixes and 27.45% of random coil. Likewise, the topology structure of SRF-TF domain for HsMADS2 protein were also consisted of one α helix. However, 58% of alpha helixes, 28% of random coil and 14% of extended strands were discovered in the secondary structure SRF-TF of HsMADS2 protein (Fig. 5).

Discussion

The significant similarities in the protein sequences of HsMADS1 and HsMADS2 with MADS-17 of *G. hirsutum* (accession no: AEJ76841.1) and MADS-13 of *G. hirsutum* (accession no: FJ409870.1), respectively were not surprising as both plant species and *H. sabdariffa* belong to the same family of Malvaceae (Silva and Figueira, 2005). In fact, HsMADS1 protein was also found significantly homologous to MADS-box protein of *Carica papaya* (ACD39982.1) and *Citrus sinensis* (XP_006482430.1), which are all core eudicots and belong to the rosids but different family from *H. sabdariffa* (Viaene et al. 2010). According to Charon et al. (2012), genome duplication is responsible for the evolution and deviation MADS-box gene from its ancestor. Thus, the duplication events on MADS-box genes supported the high variation in their genome sequences to non-family members of other plant species.

The 44.5% of similarity found in the deduced amino acids between *HsMADS1* and *HsMADS2* was a less satisfying score and the two conserved motifs identified in the C region of both *HsMADS1* and *HsMADS2* suggests the possibilities that these two may be different genes. The differences in the HsMADS1 and HsMADS2 protein sequences and structures might be due to the nature of SEPALLATA and AGAMOUS genes in MADS-box family which tend to undergo gene duplication events. This was proven by Zahn et al. (2005) when they isolated nine new *SEP* genes from various plants. Phylogenetic analyses of these *SEP* sequences showed that several gene duplication events occurred during the evolution of this gene subfamily, providing potential opportunities for functional divergence. The timing of the first *SEP* duplication approximately coincided with the duplications in the DEFICIENS/GLOBOSA and AGAMOUS MADS-box subfamilies. These coincide evolution was suspected to lead to genome-wide duplication in the ancestor of extant angiosperms or multiple independent duplication events. The evolution provided new possibilities of genetic interactions between these MADS-box genes that may have been important in the origin of the flower. The distinct and highly conserved MADS and K-box domains predicted in both *HsMADS1* and *HsMADS2* may be accountable for different functions. The MADS domain is essential for DNA binding as dimers that serve as the primary DNA-binding element (Molkentin et al. 1996). According to Yang et al. (2003), the MADS domain is commonly found associated with K-box domain, which happened to be a promoter for dimerisation. In contrast to the MADS and K-box domains, the I and C domains in HsMADS1 and HsMADS2 showed more sequence variations in the two isolated genes. This finding is supported by the studies of Yang et al. (2003) and Alvarez-Buylla et al. (2000). Yang et al. (2003) proclaimed that I domain is a weakly conserved domain which play a role in the determination for the selective formation of DNA-binding dimers. On the other hand, according to Alvarez-Buylla et al. (2000), the carboxyl- terminal (C) domain is poorly-conserved domain in the MIKC MADS-box gene.

Through BLAST P, BLAST X and multiple sequences alignments of *HsMADS1* and *HsMADS2* with its homologous sequences from other species, both genes predicted to be in the same class E of MADS-box gene corresponding to the ABCDE Model. HsMADS1 protein shared significant similarity with the MADS-box protein (MADS1) isolated from *C. papaya* (ACD39982.1). Lee et al. (2014) stated that, apart from being classified as a type II MIKC MADS-box gene, MADS-box gene isolated from *C. papaya* also happened to belong in the class E of MADS-box gene. On the other hand, HsMADS2 protein was found to be closely related with MADS-box gene isolated from kenaf (HQ315826). Through phylogenetic analysis, Chen et al. (2012) suggested the MADS-box gene of kenaf belong to *AGL-6* subfamily in the class E of MADS-box gene. Based on the conserved motifs predicted in HsMADS1 protein (Fig. 1), it may be related to the SEPALLATA (*SEP*) subfamily of MADS-box due to the presence of the two conserved motifs which are CNPTLQIGY (internal) and GFIPGWML (terminal) in the C domain regions. These two motifs are strong indicators of *SEP* family gene as revealed by Tsafsaris et al. (2007) from the study on motifs in different family members of MADS-box genes from *C. sativus* L. The members of the *SEP* MADS-box subfamily are required for specifying the "floral state" by contributing to floral organ and meristem identity. In addition, Paolacci et al. (2007) also suggested that *SEP* belongs to the class E of floral homeotic gene. The gene family members expressed low level of transcript in vegetative tissue like the coleoptile, leaf and stem but highly expressed in lemma, palae and moderate in stamen, thus proposed that the function of this gene in floral organ identity. Therefore, it is possible that the *HsMADS1* isolated in this study may involve in determining in the floral organ in roselle specify in young or mature calyx tissue. Two unique motifs which are CDHEPVLQIGY (internal) and FIHGWVI (terminal) were predicted in the C-terminal region of the deduced amino acids of *HsMADS2* which linked the gene with AGAMOUS-like 6 (*AGL6*) subfamily from MADS-box gene. This assumption is supported by the study of Tsafsaris et al. (2007) where they discovered the same conserved motifs in *AGL6* family members of MADS-box genes from *C. sativus* L. Besides that, the deduced amino acids of *HsMADS2* is also highly resemblance to the MADS-box protein deduced from *Hibiscus cannabinus* (*HcMADS-box*) isolated by Chen et al. (2012). *HcMADS-box* protein was found to be highly homologous with the *AGL2*/*SEP* from *Arabidopsis* which happened to be in the class E of MADS-box as well. Genes belong to the class E of MADS-box are mainly involved in determining the formation of floral organs and crucial for the specification of sepals, petals, stamens, carpels, and ovules in angiosperms (Zhang et al. 2009; Pelaz et al. 2001). The necessity of the genes from this class E of MADS-box might be critical for flower organ developmental as Lee et al. (2014) found that the three genes from class E of MADS-box (*CpMADS1, CpMADS2* and *CpMADS3*) isolated from *C. papaya* were expressed since early development of the flower bud. The expression of the

MADS-box genes increased significantly especially before the maturation of the buds. The different motifs present at the C-terminal domains of HsMADS1 and HsMADS2 proteins suggested that both proteins are from different subfamilies of MADS-box genes. This prediction is in agreement with the study of Vandenbussche et al. (2003), they stated that proteins of the different subfamilies in MADS-box genes were characterized by distinct sequence motifs in their C-terminal domains. Frameshift mutations that occured in the evolution of MADS-box genes may have contributed to the diversification of the MADS-box genes. The clustering of the two *HsMADS* genes into two different major clades in phylogenetic analysis, where one clade consisted of members from the *SEP* subfamily and the other clade consisted of *AGL* subfamily (Fig. 3), suggested that both *HsMADS1* and *HsMADS2* possibly involved in different functions. *HsMADS1* may play a role in the formation of the bud and flower of roselle. This prediction was supported by Lai et al. (2011) who discovered that *SEP* gene from *G. hirsutum* (*GhSEP1*) involved in the expression of the *SEP* genes in stem, leaf, bud and flower. Meanwhile *HsMADS2* may be a regulator for axillary meristem formation in roselle and also linked with *AGL6* expression in the calyx during late flowering. The assumption was supported by the phylogenetic tree association as *AGL6* of *A. thaliana* (*AtAGL6*) isolated by Huang et al. (2012) was orthologous with *HsMADS2*. The *AGL6* regulated the flowering time loci and interacted with REDUCED SHOOT BRANCHING2 (*RSB2*) to control the axillary meristem formation in *A. thaliana*. Besides that, they have also found that *AGL6* was highly expressed to suppress *RSB* phenotype during late flowering stages for stem branching purpose. Apparently, *AGL6* facilitated the formation of the axillary meristem in *A. thaliana* during the reproductive phase. The primary structure of *HsMADS1* showed resemblance to SEPALLATA 1 (*SEP 1*) $MIKC^c$ – MADS-box gene that was isolated by Riese et al. (2005) from the *Physchomitrella patens* (moss). Both *HsMADS1* and *SEP1* of *P. patens* were similar in term of their exon-intron organizations, which were composed of seven exons and six introns with similar possession of M, I, K and C domains in the sequence (Fig. 2). Apart from that, the exon-intron organization of *HsMADS1* gene sequence also shared similarity with *SEP* MADS-box gene isolated by Zhang et al. (2009) from *Populus deltoides*. The strong resemblence of the primary structures of *HsMADS1* gene with *SEP* MADS-box genes of other species further strengthen the assumption that *HsMADS1* belongs to the *SEP* subfamily of MADS-box genes. Different from *HsMADS1*, the primary structure of *HsMADS2* showed more nucleotides similarity with MADS-box gene from the *AGL6* subfamily. It was reported by Chen et al. (2012) that the full length gene of MADS-box gene from *Hibiscus cannabinus* (*HcMADS-box*) gene possesed eight exons and seven introns, which is similar with the *HsMADS2* gene. Besides belonging to the E-class MADS box gene, MADS-box gene of *H. cannabinus (HcMADS-box)* isolated by Chen et al. (2012) was also predicted to belong to the Agamous–Like 6 family (*AGL6*). Li et al. (2010) verified that the gene structure of MADS box gene from *Oryza sativa* (*OSMADS6*) consisted of eight exons and seven introns as well. Similar to *HsMADS2*, *HcMADS-box* and *OSMADS6* genes also have the longest first exon among all other exons that signaturely encoded the MADS domain. In this study, Serum Response Factor-Transcription Factor (SRF-TF) domain were predicted in the MADS domain of *HsMADS1* and *HsMADS2*. Pellegrini et al. (1995) and Norman et al. (1988) stated that the human serum response factor (SRF) is a transcription factor belonging to the MADS domain protein family with members characterized from the plant and animal kingdoms. Thus, the presence of SRF-TF domain in *HsMADS1* and *HsMADS2* were not surprising. SRF-TF domain is involved in DNA-binding and dimerisation of MADS domain (Norman et al. 1988; Huang et al. 1996). The SRF-TF domain predicted in HsMADS1 and HsMADS2 is a signature for MADS domain. The SRF-TF domains in HsMADS1 and HsMADS2, which consisted of only α helix are similar with the SRF- TF structure studied by Pellegrini et al. (1995). They discovered that the structure of SRF-TF were consisted of two α helices, without β strand. It is also in agreement with the findings of Sinha and Sengupta (2013) where they discovered that the SIMADS RIN protein from tomato was also a helix-rich protein with 54.13% of alpha helix found in the protein they studied. Pellegrini et al. (1995) highlighted the importance of SRF in the N-terminal α-helix of the MADS-box in DNA binding and providing dimerisation interface for dimer formation in MADS-box genes. The topology of β- α- β- α- α- α pattern from the secondary structure of HsMADS1 and HsMADS2 were significantly matched with the crystal structure of MADS-box/Myocyte Enhancer Factor-2 (MEF2) domain from *Homo sapiens* (PDB ID: 1N6J). 1N6J consisted of a α- β- β- α- α pattern (Han et al. 2003). In addition, the MEF2 domain from silkworm investigated by Ling et al. (2008) also has the similar topology structure like 1N6J. The first β- α- β in HsMADS1 and HsMADS2 probably representing the MADS domain. According to Huang et al. (2000), the N-terminal tail, first helix (α1), first strands (β1) and second strand (β2) together formed the MADS-box core domain. The MADS domain mediates DNA recognition and dimerisation in a manner similar to that observed in the MEF2A/DNA complexes. Han et al. (2003) revealed that MADS-box/MEF2S domain of human MEF2B bound to a motif of the transcriptional co-repressor Cabin1 and DNA. The crystal structure was a stably folded MEF2S domain on the surface of the MADS-box. They stated that Cabin1 structure in MEF2 domain acts as co-repressor and adopting an amphipathic α-helix to bind a hydrophobic groove on the MEF2S domain, forming a triple-helical interaction. Thus, we hypothesised that the MEF2 associated with MADS domain in HsMADS1 and HsMADS2 may possibly involve in repressing or activating the transcription factor for DNA binding affinity in roselle by the association with co-repressor at their respective C-terminals with dependency with Calcium (Ca^{2+}).

Materials and Methods

Plant samples

Seeds of *Hibiscus sabdariffa* var. UMKL were obtained from the Department of Agriculture Terengganu, Malaysia. The type of planting medium used was the mixture of humus and sandy soils in the ratio of 3:1. Fresh calyces for two developmental stages; young and mature were collected from own cultivated roselles after 53 and 61 days of planting respectively (Fig. 6).

RNA isolation

Fresh young and mature calyx tissues collected from the *H. sabdariffa* were used for RNA isolation using the RNeasy Plant Mini Kit (Qiagen, Germany) according to its standard protocol.

CDS isolation with Rapid Amplification of cDNA Ends (RACE)-PCR

Gene specific primers (GSP) for CDS amplification were designed based on the consensus nucleotides sequences of six MADS-box genes retrieved from NCBI following multiple sequence alignment. The first strand of *HsMADS1* and *HsMADS2* CDS were amplified with 3' RACE PCR using the following RACE PCR primer, MADS3R: 5'-ATGGGAAGAGGAAGAGTAGAGCTG-3'. The first strand cDNA of HsMADS1 and HsMADS2 were synthesized with Touchdown PCR program 1, according to the recommended protocol in SMARTer RACE cDNA Amplification Kit for GSP with Tm > 70 ºC (Clontech, USA).

DNA extraction and polysaccharides removal

Fresh young and mature calyces of *H. sabdariffa* were used for DNA isolation by using the conventional DNA extraction method from Doyle and Doyle (1990) protocol with a modification. An additional step was added after the DNA extraction to reduce the polysaccharides in the DNA sample. Briefly, 500 µl of DNA was treated with 250 ul 5.0 M of Sodium chloride (NaCl) in a 1.5 ml microcentrifuge tube and kept on ice for 30 minutes before be centrifuging 17,000 xg for 10 minutes. The supernatant containing DNA was collected and transferred into a new 1.5 ml microcentrifuge tubes, and was precipitated with 1 ml of 100% absolute ethanol for one hour at -20 ºC. The tube containing the DNA was then centrifuged at 2,278 xg for 10 minutes. The supernatant was discarded and the DNA pellet was dissolved in 500 µl ddH_2O for six hours at 4 ºC after being air-dried. DNA was further purified using phenol and chloroform-isoamyl alcohol method.

Isolation of the intronic regions of HsMADS1 and HsMADS2

HsMADS1 and *HsMADS2* genes were amplified by PCR using the following forward-reverse primers pairs, HsMADS1F1**:**5'-GGGATGCTGAGGTTGCTCT-3', HsMADS1R1: 5'-CCAGGGATAAAGCCATTGAC-3' and HsMADS2F1: 5'-ATGGGAAGAGGAAGAGTAGAGTTG-3', HsMADS2R1: 5'-TGTGAGTTGGAGACGGTTCA-3', respectively. The PCR reactions were carried out in 20 µl reactions volume in different tube for *HsMADS1* and *HsMADS2* genes amplification. The two tubes of 20 µl reactions containing 27ng/µl (DNA of young calyx) and 40ng/µl (DNA of mature calyx), respectively. The final concentration of the other PCR components in 20 µl reaction were; 0.125 µM for each forward primer and reverse primer,1x Buffer, 1.875 mM of Magnesium Chloride ($MgCl_2$), 0.2 mM of dNTPs and 1.25 unit of Taq polymerase. PCR amplifications were conducted in PTC-200 Thermal Cycler PCR (MJ Research, Germany). The cycling profile started with a cycle of pre-denaturation at 95ºC for 1 min 30s, following by 35 cycles of [95ºC/30s, 52ºC (for DNA isolated from young calyx); 54.1 ºC (for DNA isolated from mature calyx)/30s, 72ºC/2 min] and ended with a cycle of final extension at 72ºC for 7 min. The amplicons were purified using QIAquick Gel Extraction kit (QIAGEN, Germany) and cloned using pGEM-T Easy Vector System (Promega, USA). The cloned PCR products were purified using Wizard Plus SV Miniprep DNA purification kit (Promega, USA) and sent for sequencing using ABI platform.

In-silico analysis

The full length coding (CDS) and gene sequences of *HsMADS1* and *HsMADS2* were analysed using BLAST analyses (Altschul et al. 1997). Homologous nucleotide sequences that correspond to these two target sequences were aligned using Clustal W Version 2.0 to identified regions that are conserved. The Open Reading Frame predicted from the CDS of *HsMADS1* and *HsMADS2* were further deduced into amino acid sequences using Emboss Protein Translation Tools (Rice et al. 2000) and analysed by Interproscan (Hunter et al. 2009), Pfam (Finn et al. 2014) and BLAST P for domain prediction. Poly-A tails were predicted using GENSCAN Web Server (Burge and Karlin, 1997). You should explain a bit how the assembly was done since you mentioned about the assembled sequences in your result under cloning of CDS.

Phylogenetic tree

Phylogenetic analyses between HsMADs1 and HsMADs2 with 37 sequences of MADS genes from other species were performed using neighbour-joining (NJ) method. The bootstrap values of the phylogenetic trees were derived from 2,000 replicates run. The Genbank accession numbers of the 37 amino acid sequences used are; *GaSEP2* (KHG17252.1; *G. arboretum*), *GhMADS17* (AEJ76841.1; *G. hirsutum*), *GhMADS62* (AGW23364.1; *G. hirsutum*), *TcMADS* (XP_007032865.1; *Theobroma cacao*), *CsSEP1* (XP_006482430.1; *C. sinensis*), *CpMADS1* (ACD39982.1; *C. papaya*), *CaSEP1* (XP_004507288.1; *Cicer arietinum*), *MtMADS* (KEH25149.1; *Medicago truncatula*), *GmSEP1.1* (XP_003544012.1; *Glycine max*), *JcSEP1* (XP_012088064.1; *Jatropha curcas*), *GsSEP2* (KHN29767.1; *Glycine soja*), *PmSEP1* (XP_008230292.1; *Prunus mume*), *CsSEP2* (NP_001267667.1; *Cucumis sativus*), *PtMADS4* (ABG78619.1; *Populus tomentosa*), *PaMADS4* (AFM30904.1; *Prunus avium*), *MiSEP1* (ADX97328.1; *Mangifera indica*), *NnSEP1* (XP_010257958.1; *Nelumbo nucifera*), *GhMADS13* (ACJ26768.1; *G. hirsutum*), *GhMADS45* (AGW23347.1; *G. hirsutum*), *GaAGL6* (KHF99411.1; *G. arboreum*), *TcAGL6* (XP_007051983.1; *T. cacao*), *VvMADS3* (NP_001268111.1; *Vitis vinifera*), *HcMADS* (ADZ98838.1; *H. cannabinus*), *JcAGL6* (XP_012083518.1; *Jatropha curcas*), *RcMADS* (XP_002511765.1; *Ricinus communis*), *GhMADS46* (AGW23348.1; *G. hirsutum*), *MdAGL6* (NP_001280892.1; *Malus domestica*), *PbAGL6* (XP_009375842.1; *Pyrus x bretschneideri*), *CusAGL6* (XP_011656687.1; *Cucumis sativus*), *CaAGL6* (XP_004492666.1; *Cicer arietinum*), *NcMADS6* (XP_010272608.1; *N. nucifera*), *CasAGL6* (XP_010518186.1; *Camelina sativa*), *AtAGL6* (AFP23750.1; *A. thaliana*), *ApSEP1* (BAC66964.1; *Agapanthus praecox*), *CmAGL6* (XP_008460143.1; *Cucumis melo*), *WmAGL6* (AGV28075.1; *Welwitschia mirabilis*), *AtAP1* (NP_177074.1 ; *A. thaliana*).

Comparative modelling of 3D structures

The templates for HsMADS1 and HsMADS2 proteins were selected based on PSI-BLAST. The crystal structure of Keratin-like Domain from MADS transcription factor Sepallata 3 of *A. thaliana* (PDB accession: 4OX0) was the most significant template with the longest region of similarity with HsMADS1 followed by 1EGW, 3KOV, 3MU6 and 1TQE (Supplementary Table 3). The most significant template for the HsMADS2 was found with crystal structure

of Mef2a core bound to *H. sapiens* (PDB accession: 1EGW). Apart from that, HsMADS2 was predicted to be matched with four other templates sequences which are 4OX0, 3KOV, 1TQE and 1C7U (Supplementary Table 4) The three dimensional (3D) homology structures of HsMADS1 and HsMADS2 proteins were constructed by MODELLER version 9.13 (Sali and Blundell, 1993). The stereochemistry errors of the HsMADS1 and HsMADS2 protein models were validated by RAMPAGE: Ramachandran Plot (Lovell et al. 2003) and ProSA (Sippl, 1993; Wiederstein and Sippl, 2007).

Conclusion

HsMADS1 and *HsMADS2* were predicted to have different structures and functions from each other. Both genes might carry a significant role at different developmental stages of calyx stages in roselle, respectively. Nevertheless, further characterizations need to be carried out in order to confirm the functions of both *HsMADS1* and *HsMADS2* in the calyx of roselle as well as studying the mechanism of the genes molecular functions.

Acknowledgements

The study was funded by Research University Grant Scheme (RUGS, Project Number: 05-02-12-2173RU).

References

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, and Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25: 3389-3402.

Alvarez-Buylla ER, Pelaz S, Liljegren SJ, Gold SE, Burgeff C, Ditta, GS, de Pouplana LR, Marti´nez-Castilla L, Yanofsky MF (2000) An ancestral MADS-box gene duplication occurred before the divergence of plants and animals. PNAS. 97:10.

Burge C, Karlin S (1997) Prediction of complete gene structures in human genomic DNA. J Mol Bio. 268: 78-94.

Charon C, Bruggeman Q, Thareau V, Henry Y (2012) Gene duplication within the Green Lineage: the case of TEL genes. J Exp Bot. 63: 5061–5077.

Chen P, Li R, Liao J, Zhou, R (2012) Cloning, expression and characterization of a novel MADS-box gene from kenaf (*Hibiscus cannabinus* L.). J Anim Plant Sci. 13(1): 1714-1724.

Da-Costa-Rocha I, Bonnlaender B, Sievers H, Pischel I, Heinrich, M (2014) *Hibiscus sabdariffa* L. – A phytochemical and pharmacological review. Food Chem. 165: 424–443.

de Oliveira RR, Cesarino I, Mazzafera P, Dornelas MC (2014) Flower development in *Coffea arabica* L.: new insights into MADS-box genes. Plant Reprod. 27: 79-94.

Doyle JJ, Doyle JL (1990) Isolation of plant DNA from fresh tissue. Focus 12: 11-15.

Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L., Mistry J, Sonnhammer E. L. L., Tate, J and Punta M (2014) Pfam: the protein families database. Nucleic Acids Res. 42: 222-230.

Han A, Pan F, Stroud JC, Youn HD, Liu JO, Chen L (2003) Sequence-specific recruitment of transcriptional co-repressor Cabin1 by myocyte enhancer factor-2. Nature 422: 730-734.

Huang H, Tudor M, Su T, Zhang Y, Hu Y, Maa H (1996) DNA Binding Properties of Two Arabidopsis MADS Domain Proteins: Binding Consensus and Dimer Formation. The Plant Cell 8: 81-94.

Huang K, Louis JM, Donaldson L, Lim FL, Sharrocks AD, Clore GM (2000) Solution structure of the MEF2A-DNA complex: structural basis for the modulation of DNA bending and specificity by MADS-box transcription factors. Embo J. 19:2615–2628.

Huang X, Effgen S, Meyer RC, Theres K, Koornneef M (2012) Epistatic Natural Allelic Variation Reveals a Function of AGAMOUS-LIKE 6 in Axillary Bud Formation in *Arabidopsis*. The Plant Cell 24(6): 2364–2379.

Hunter H, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Das U, Daugherty L, Duquenne L, et al (2009) InterPro: the integrative protein signature database. Nucleic Acids Res. 37(1): 211-215.

Lai D, Li H, Fan S, Song M, Pang C, Wei H, Liu J, Wu D, Gong W, Yu S (2011). Generation of ESTs for Flowering Gene Discovery and SSR Marker Development in Upland Cotton. PLoS One. 6(12): e28676.

Lee MJ, Yang WJ, Chiu CT, Chen JJ, Chen FC, Chang LS (2014) Isolation and characterization of the papaya MADS-box E-class genes, *CpMADS1* and *CpMADS3*, and a TM6 lineage gene *CpMADS2*. Genet Mol Res.13(3): 5299-5312.

Li HF, Liang WQ, Jia RD, Yin CS, Zong J, Kong HZ, Zhang DB (2010) The *AGL6*-like gene *(OsMADS6)* regulates floral organ and meristem identities in rice. Cell Res. 20: 299–313.

Ling QZ, Yu M, Zhang JQ, Chu LH, Wei ZJ (2008) Molecular characters and expression analysis of a new isoform of the myocyte enhancer factor 2 gene from the silkworm, *Bombyx mori*. Afr J Biotechnol. 7(20): 3587-3592.

Lovell SC, Davis IW, Arendall, WB, de Bakker PI, Word JM, Prisant, MG, Richardson JS, Richardson DC (2003) Structure validation by C alpha geometry: phi, psi and C beta deviation. Proteins 50(3): 437-450.

Mgaya-Kilima B, Remberg SF, Chove BE, Wicklund T (2015) Physiochemical and antioxidant properties of roselle-mango juice blends; effects of packaging material, storage temperature and time. Food Sci Nutr. 3(2): 100–109.

Mohamad O, Ramadan G, Halimaton-Saadiah O, Noor-Baiti AA, Marlina MM, Nurul Rahainah CM, Aulia-Rani A, Elfi K, Nur-Syakireen I, Salwa AS, Nur-Samahah MZ, Zainal M, Mamot S, Jalifah L, Aminah A, Golam F, Ahmad-Bachtiar B, Mohd- Nazir B, Mohd-Zulkifli MZ., Abd. Rahman M, Zainal AA (2009) Development of three new varieties of roselle by mutation breeding. Paper presented at Proceedings of the 8th Malaysia Congress on Genetics, Genting Highlands, Malaysia, 4-6 August, 2009.

Molkentin JD, Black BL, Martin JF, Olson EN (1996) Mutational Analysis of the DNA Binding, Dimerization and Transcriptional Activation Domains of MEF2C. Mol Cell Biol. 16(6): 2627–2636.

Norman C, Runswick M, Pollock R, Treisman R (1988) Isolation and properties of cDNA clones encoding SRF, a transcription factor that binds to the c-fos serum response element. Cell 55 : 989–1003.

Osman M, Golam F, Saberi S, Abdul-Majid N, Nagoor NH, Zulqarnain, M (2011) Morpho-agronomic analysis of three roselle (*Hibiscus sabdariffa* L.) mutants in tropical Malaysia. Aust J Crop Sci. 5(10):1150-1156.

Paolacci AR, Tanzarella OA, Porceddu E, Varotto S, and Ciaffi M (2007) Molecular and phylogenetic analysis of MADS-box genes of MIKC type and chromosome location of *SEP*-like genes in wheat (*Triticum aestivum* L.). Mol Genet Genomics. 278(6): 689-708.

Pelaz S, Gustafson-Brown C, Kohlami SE, Crosby WL, Yanofsky MF (2001). APETALA1 and SEPALLATA3 interact to promote flower development. Plant J. 26**:** 385–394.

Pellegrini L, Tan S, Richmond, TJ (1995) Structure of serum response factor core bound to DNA. Nature 376: 490–498.

Rice P, Longden I, Bleasby A (2000). EMBOSS: The European Molecular Biology Open Software Suite. Trends Genet. 16: 276-7.

Riese, M, Faigl, W, Quodt V, Verelst, W, Matthes A, Saedler H, Münster T (2005) Isolation and characterization of new MIKC*-type MADS-box genes from the moss Physcomitrella patens. Plant Biol. 7: 307–314.

Sali A, Blundell, TL (1993) Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol. 234: 779-815.

Silva CRS, Figueira A (2005) Phylogenetic analysis of *Theobroma* (Sterculiaceae) based on Kunitz-like trypsin inhibitor sequences. Plant Syst Evol. 250: 93–104.

Sinha SK, Sengupta DN (2013) Homology Modeling of a Fruit Ripening Specific Plant MADS–box Factor**.** Am J Biochem Mol Biol. 3(2): 188-201.

Sippl MJ (1993) Recognition of Errors in Three-Dimensional Structures of Proteins. Proteins 17: 355-362.

Song JJ, Ma W, Tang YJ, Chen ZY, Liao JP (2010) Isolation and characterization of three MADS-box genes from *Alpinia hainanensis* (Zingiberaceae). Plant Mol Biol Rep. 28: 264-276.

Tsaftaris AS, Pasentsis K, Kalivas A, Polidoros AN (2005) The Family of MADS – Box Genes Controlling Flower Development in Crocus (*Crocus sativus* L.). Paper presented at 13th International Congress on Genes, Gene Families and Isozymes – ICGGFI, Shanghai, China, 17-21 September 2005

Tsaftaris AS, Polidoros AN, Pasentsis K (2007) Cloning, Structural Characterization, and Phylogenetic Analysis of Flower MADS-Box Genes from Crocus (*Crocus sativus* L.). Special Issue: Evolution of MADS-Box Genes in Monocots. Sci World J. 7: 1047–1062.

Vandenbussche M, Theissen G, Van de Peer Y, Gerats T (2003) Structural diversification and neo-functionalization during floral MADS-box gene evolution by C-terminal frameshift mutations. Nucleic Acids Res. 31: 4401–4409.

Viaene T, Vekemans D, Becker A, Melzer S, Geuten K (2010) Expression divergence of the *AGL6* MADS domain transcription factor lineage after a core eudicot duplication suggests functional diversification. BMC Plant Biol. 10:148.

Wiederstein M, Sippl MJ (2007) ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. Nucleic Acids Res. 35: 407-410.

Yang Z, Fanning L, Jack T (2003) The K domain mediates heterodimerization of the Arabidopsis floral organ identity proteins, APETALA3 and PISTILLATA. The Plant J. 33: 47–59.

Zahn LM, Kong H, Leebens-Mack JH, Kim S, Soltis PS, Landherr LL, Ma H (2005) The Evolution of the *SEPALLATA* Subfamily of MADS-Box Genes: A Preangiosperm Origin With Multiple Duplications Throughout Angiosperm History. Genetics 169(4): 2209–2223.

Zhao TJ, Zhao SY, Chen HM, Zhao QZ, Hu ZM, Hou BK, Xia GM (2006) Transgenic wheat progeny resistant to powdery mildew generated by *Agrobacterium inoculum* to the basal portion of wheat seedling. Plant Cell Rep. 25: 1199–1204.

Zhang B, Zhang Z, Li H, Zhou X, Su X (2009). Cloning and expression analysis of an E-class MADS-box gene from *Populus deltoide.* Afr J Biotechnol. 8(19): 4789-4796.