Pak. J. Bot., 49(3): 1113-1126, 2017.

GENOME-WIDE IDENTIFICATION AND COMPARATIVE ANALYSIS OF SQUAMOSA-PROMOTER BINDING PROTEINS (SBP) TRANSCRIPTION FACTOR FAMILY IN *GOSSYPIUM RAIMONDII* AND *ARABIDOPSIS THALIANA*

MUHAMMAD AMJAD ALI[1,2§], KHUSH BAKHAT ALIA[5§], RANA MUHAMMAD ATIF[3,4], IJAZ RASUL[5], HABIB ULLAH NADEEM[5], AMMARA SHAHID[5] AND FARRUKH AZEEM[*5]

[1]*Department of Plant Pathology, University of Agriculture, 38040 Faisalabad, Pakistan*
[2]*Centre of Agricultural Biochemistry & Biotechnology, University of Agriculture, 38040 Faisalabad, Pakistan.*
[3]*US-Pakistan Centre for Advanced Studies in Agriculture and Food Security, University of Agriculture, 38040 Faisalabad, Pakistan*
[4]*Department of Plant Breeding & Genetics, University of Agriculture, 38040 Faisalabad, Pakistan*
[5]*Department of Bioinformatics & Biotechnology, Government College University, 38000 Faisalabad*
[§] *These authors have equally contributed to the work*
[*]*Corresponding author's email: azeuaf@hotmail.com*

Abstract

SQUAMOSA-Promoter Binding Proteins (SBP) are class of transcription factors that play vital role in regulation of plant tissue growth and development. The genes encoding these proteins have not yet been identified in diploid cotton. Thus here, a comprehensive genome wide analysis of SBP genes/proteins was carried out to identify the genes encoding SBP proteins in *Gossypium raimondii* and *Arabidopsis thaliana*. We identified 17 SBP genes from *Arabidopsis thaliana* genome and 30 SBP genes from *Gossypium raimondii*. Chromosome localization studies revealed the uneven distribution of SBP encoding genes both in the genomes of *A. thaliana* and *G. raimondii*. In cotton, five SBP genes were located on chromosome no. 2, while no gene was found on chromosome 9. In *A. thaliana*, maximum seven SBP genes were identified on chromosome 9, while chromosome 4 did not have any SBP gene. Thus, the SBP gene family might have expanded as a result of segmental as well as tandem duplications in these species. The comparative phylogenetic analysis of Arabidopsis and cotton SBPs revealed the presence of eight groups. The gene structure analysis of SBP encoding genes revealed the presence of one to eleven inrons in both Arabidopsis and *G. raimondii*. The proteins sharing the same phyletic group mostly demonstrated the similar intron-exon occurrence pattern; and share the common conserved domains. The SBP DNA-binding domain shared 24 absolutely conserved residues in Arabidopsis. The present study can serve as a base for the functional characterization of SBP gene family in *Gossypium raimondii*.

Key words: SQUAMOSA-Promoter binding proteins, Transcription factors, Diploid cotton, Phylogenetic analysis, *In silico*.

Introduction

Transcription factors (TFs) are DNA binding proteins that control gene expression in living organisms, as in plants these hold a strategic role at the level of gene regulation (Wang *et al*., 2009). In plants several families of transcription factors are identified which mediates plant responses right from seed germination to plant maturity (Noguero *et al*. 2013). The genome of Arabidopsis alone contains more than 25 major families of transcription factors (Riechmann & Rateliffe 2000). One of these families comprises a DNA binding domain known as the SQUAMOSA PROMOTER BINDING PROTEIN (SBP) domain which is encoded by the SBP-box. The member of this family are called SBP transcription factors.

A numeral of biochemical and functional studies revealed that the SBP-box gene show vital role in regulation of plant tissue growth and development. The proteins belonging to SPB family are distinct by a extremely sealed area of 76 amino acids called the SBP domain (Preston & Hileman, 2013). The SBP domain is complicated and contains sequence-specific DNA binding to a consensus-binding site comprising a GTAC core motif (Preston & Hileman, 2013). The extremely conserved 76 amino acids residues SBP domain binds precisely to associated motifs in the *Antirrhinum majus* SQUA promoter and the orthologus *Arabidopsis thaliana* AP1 promoter (Birkenbihl *et al*., 2005). The arrangement bases for this sequence specific binding arrangements formed by the organization of two zinc ions by conserved cysteine and histidine residues (Cardon *et al*., 1999). SPL genes are present in all green plants counting single celled, mosses, green algae, angiosperms and gymnosperms, and (Preston and Hileman, 2013). In recent years, SBP genes are recognized in various plants for example a novel BPSPL1, SBP-box gene from silver birch (*Betula pendula*) binds with a *cis* element of BpMADS5 (Lannenpaa *et al*., 2004).

SBP-box is also a feature characteristic of the *Arabidopsis* SQUAMOSA PROMOTER BINDING PROTEIN–LIKE (SPL) gene family (Wang *et al*., 2009). A gene belonging to this gene family, *SPL* is found to play role in initial stages of microsporogenesis and megasporogenesis and as well as the growth of ordinary plant architecture (Cardon *et al*., 1997). SPL gene family was first identified in *Antirrhinum majus*. SPL genes are closely related to AmSBP1 and AmSBP2 to bind to the promoter of the floral meristem unique gene of SQUAMOSA (Klein *et al*., 1996). The members of gene families containing SBP box have been extensively reviewed recently (Preston & Hileman, 2013).

SBP transcription factors have been continued the effort of *In silico* studies in numerous plant species. In the latest era, owing to the improvements in sequencing techniques and data analysis, the genomes of numerous

crop species have been sequenced. These genomes have been topic of bioinformatics analysis for various transcription factor gene families (Noguero *et al*. 2013; Khan *et al*. 2016). There are several reports of bioinformatics studies of SBP transcription factors in different plant species (Guo *et al*., 2008; Yamasaki *et al*., 2004).

In this study, we have focused on the genome wide analysis (Imran & Liu, 2016; Watanabe & Khan, 2016) of SBP genes in Arabidopsis and *Gossypium raimondii* and their genomic comparison have done to show the functional similarity among these species. Further, all the SBP genes are analyzed by using *in silico* tools. This study comprises of phylogenetic analysis, gene structure analysis, multiple sequence analysis, promoter analysis, synteny analysis and chromosomal mapping of SBP transcription factor in Arabidopsis and *Gossypium raimondii*. This study will pave the way for genomic comparison of the SBP transcription factors in these plant species to disclose functional homology in genomic and protein level.

Materials and Methods

Retrieval of sequence of SBP proteins in *Arabidopsis thaliana* and *Gossypium raimondii*: All the protein sequences of SBP genes of *ArabidopsisThaliana* and *Gossypium Raimondii* were retrieved from plant TFDB site versio3.0.(http://planttfdb.cbi.pku.edu.cn/).The intron and exon positions of both the plant species was retrieved from Phytozome9.1 (http://www.phytozome.net/).

Phylogenetic analysis: The SBP proteins of Arabidopsis and Gossypium were subjected to multiple sequence alignment using ClustalW and on the basis of this multiple sequence a phylogenetic tree was constructed using software MEGA 6.06 version. The parameters used for multiple sequence alignment were Gap opening penalty: 10;Gap extension penalty:0.2; Residue specific penalties: on; Hydrophilic penalties: on; Gap separation distance:4; End gap separation penalty: off; Negative matrix: off; Delay divergent cutoff: 30%.

Identification of conserved motifs in SBP proteins: Motif analysis of SBP proteins of Arabidopsis and cotton was performed using MEME online software version 4.9.1 (http://meme.nbcr.net/meme/cgi-bin/meme.cgi). This motif analysis was based on certain parameters. These parameters were maximum number of motifs: 20; The Maximum motif width was between 5 to 90 residues. The occurrences of a single motif among the sequences were settled to any number of repetitions.

Chromosomal mapping of SBP genes: All the SBP genes of *Arabidopsis thaliana* and *Gossypium raimondii* were mapped on chromosomes of respective plant species. This chromosomal mapping was based on the information available at Phytozome 9.1(http://www.phytozome.net/) and NCBI (http://www.ncbi.nlm.nih.gov/) databases. Chromosomal mapping was performed in excel sheet. The SBP genes of *A. thaliana* were distributed among 5 chromosomes and all the SBP genes of *G. Raimondii* were distributed among 13 chromosomes.

Multiple sequence alignment: The SBP domain was analyzed in Arabidopsis and *G. raimondii* through multiple sequence alignment. This analysis was carried out by performing multiple sequence alignment by using online tool COBALT available at NCBI (http://www.stva.ncbi.nlm.nih.gov/tools/cobalt/re_cobalt.cgi) and Unipro UGENE. Default settings were not changed in COBALT during alignment.

Gene structure analysis: For gene structure analysis the whole genomic sequences of all the SBP genes of Arabidopsis and Gossypium were downloaded from Phytozome (http://www.phytozome.net/) Version 9.1. In *Arabidopsis thaliana*, the position and the exact numeric value of intron and exon was calculated from TAIR (https://www.arabidopsis.org/). In *Gossypium raimondii* the position and the exact numeric value of intron and exon was calculated manually. The gene structure analysis was carried out by using online software Fancy Gene version 1.4(http://bio.ieo.eu/fancygene/).

Analysis of *Cis* regulatory element in promoter sequence: The 1kb promoter sequence upstream to the start codon of all SBP genes of Arabidopsis and Gossypium were retrieved from their genome assemblies by using Phytozome version 9.1 (http://www.phytozome.net/). To identify the presence of cis-regulatory elements in the promoter sequence these sequences were subjected to PLACE (http://www.dna.affrc.go.jp/PLACE/). Place facilitates the identification of motifs in the given promoter sequences.

Synteny Analysis: The protein sequences of SBP genes of both plant species retrieved from plant TFDB site versio3.0.(http://planttfdb.cbi.pku.edu.cn/) were submitted to the online synteny analysis tool circoletto (tools.bat.infspire.org/circoletto).

Results

Phylogenetic analysis of SBP proteins in arabidopsis and cotton: The SBP proteins of Arabidopsis and Gossypium *raimondii* were subjected to multiple sequence alignment using ClustalW and a phylogenetic tree was constructed using MEGA v. 6.06 software and neighbor joining (NJ) method. Based on this phylogenetic analysis, SBP genes can be divided into 8 groups in which the *Arabidopsis thaliana* and *Gossypium raimondii* genes were clustered according to the homology between the SBP protein sequences (Fig. 1). Seven proteins of Gossypium and 3 of Arabidopsis were clustered in group1 as they show greater similarity. In 2nd group, 5 proteins of Gossypium and 2 proteins of Arabidopsis were clustered. Similarly, group3 consists of 2 proteins from Arabidopsis and one from Gossypium. Group4 consist of 2 proteins of Arabidopsis and 4 proteins of Gossypium while Group5 carry 6 proteins with 3 from both plant species. Moreover, group6 clustered 3 proteins from Arabidopsis and a couple of proteins from Gossypium. The group7 contained 5 proteins, which are only coded by *Gossypiyum raimondii* genes. In 8th cluster, 3 proteins from each plant species were present. All the relative information about the SBP genes in Arabidopsis and Gossypium is shown in Tables 1 & 2.

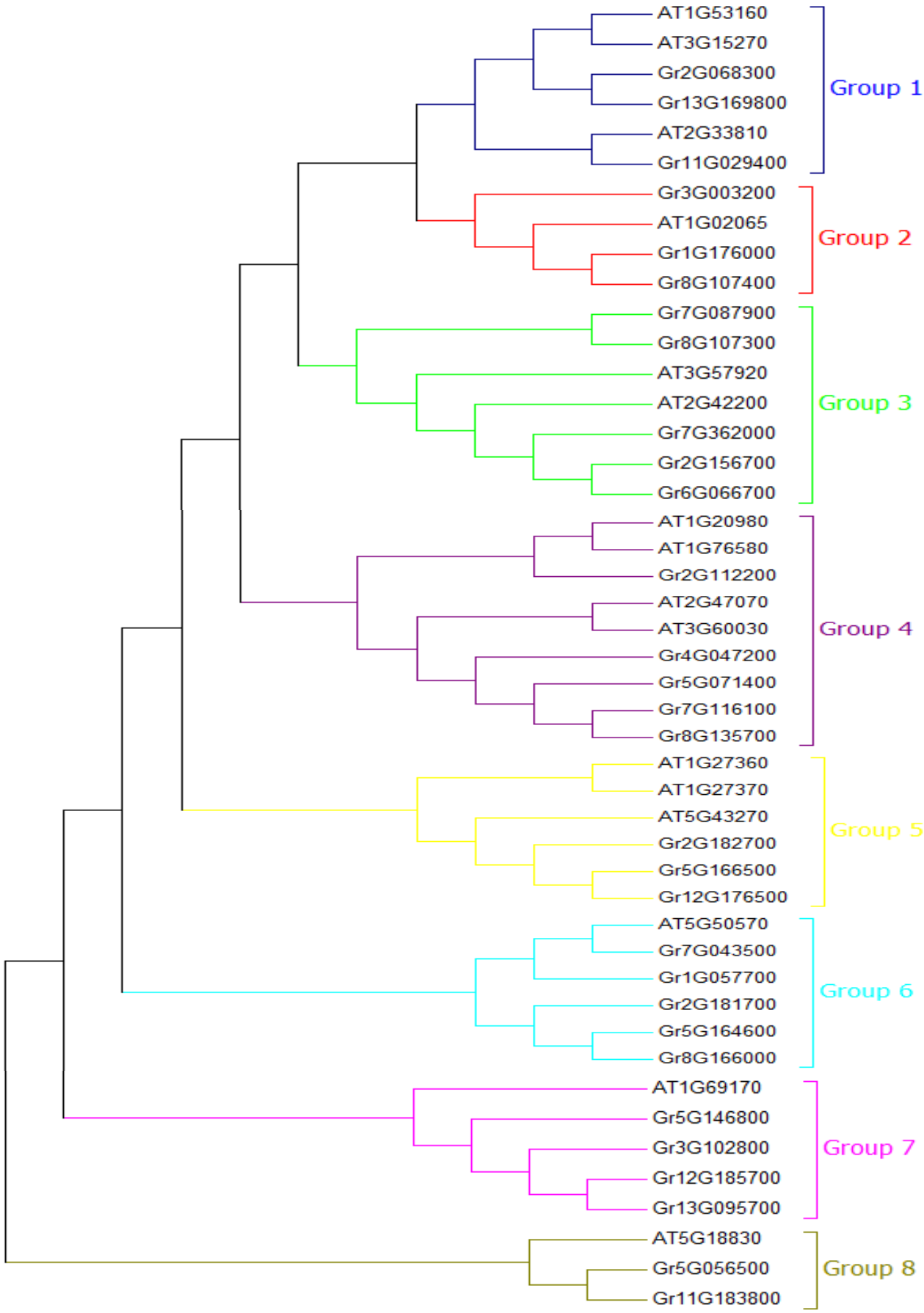

Fig. 1. Phylogenetic relationship of *Arabidopsis thaliana* and *Gossypium raimondii* SBP genes.

Table 1. SBP transcription factors identified in *Arabidopsis thaliana*.

Sr.#	Gene #	Chr #	Intron	Protein length	Gene length
1.	AT1G02065	1	2	333	2132
2.	AT1G20980	1	9	1035	4798
3.	AT1G27360	1	3	393	1899
4.	AT1G27370	1	3	396	3080
5.	AT1G53160	1	2	174	1190
6.	AT1G69170	1	3	405	2196
7.	AT1G76580	1	11	988	4684
8.	AT2G33810	2	1	131	1072
9.	AT2G42200	2	2	375	2224
10.	AT2G47070	2	9	881	4175
11.	AT3G15270	3	1	181	984
12.	AT3G57920	3	2	354	1715
13.	AT3G60030	3	9	927	3919
14.	AT5G18830	5	10	818	4578
15.	AT5G43270	5	4	419	3469
16.	AT5G50570	5	3	359	2477

Table 2. SBP transcription factors identified in *Gossypium raimondii*.

Sr.#	Gene #	Chr #	Intron	Protein length	Gene length
1.	Gr1G057700	1	3	352	2286
2.	Gr1G176000	1	3	292	2765
3.	Gr2G068300	2	1	182	1458
4.	Gr2G112200	2	11	1081	6207
5.	Gr2G156700	2	2	374	2394
6.	Gr2G181700	2	4	292	3074
7.	Gr2G182700	2	5	477	4459
8.	Gr3G003200	3	1	173	1348
9.	Gr3G102800	3	4	525	5397
10.	Gr4G047200	4	10	928	6005
11.	Gr5G056500	5	9	804	7487
12.	Gr5G071400	5	10	1040	10941
13.	Gr5G146800	5	1	201	1815
14.	Gr5G164600	5	3	309	3743
15.	Gr5G166500	5	4	474	5702
16.	Gr6G066700	6	2	391	3401
17.	Gr7G043500	7	3	423	2418
18.	Gr7G087900	7	9	420	2717
19.	Gr7G116100	7	2	985	6815
20.	Gr7G362000	7	2	347	2797
21.	Gr8G107300	8	2	355	2473
22.	Gr8G107400	8	2	335	2412
23.	Gr8G135700	8	9	987	7575
24.	Gr8G166000	8	3	302	3796
25.	Gr11G029400	11	1	195	1526
26.	Gr11G183800	11	9	802	7001
27.	Gr12G176500	12	6	478	3191
28.	Gr12G185700	12	3	379	2420
29.	Gr13G095700	13	4	472	4573
30.	Gr13G169800	13	1	181	1219

Comparison of chromosomal mapping of SBP genes in *Arabidopsis* and *Gossypium*: In Arabidopsis 17 genes were located on 4 chromosomes because chromosome number 4 does not carry any SBP transcription factor gene. Six genes were present on chromosome1, while 2nd 3rd chromosomes carried 3 SBP genes each. Moreover, 5 genes were present on chromosome 5. The distribution of SBP genes among the different chromosomes of *A. thaliana* is shown in the Fig. 2.

Gossypium raimondii contains 30 genes from SBP transcription factor family which were only present on 11 chromosomes out of total 13 genes. The distribution of SBP genes among the thirteen chromosomes of *Gossypium Raimondi* is shown in Fig. 3. Only one gene was present on both chromosome4 and chromosome6. There were two genes located on each of chromosome1, 3, 11, 12 and 13. However, chromosome2 and chromosome5 carried 5 genes each and 4 genes were present on both chromosome7 and 8. Moreover, chromosome number 9 and 10 did not show the presence of SBP genes.

Gene structure analysis of SBP genes in *Arabidopsis* and cotton: In order to generate more information regarding SBP gene family in *Arabidopsis* and diploid cotton, the structure of genes in the form of number and length of introns and exons present in the DNA sequence of SBP genes was analyzed. To some extent, exon–intron structures are conserved within different groups of SBP transcription factors as shown in Fig. 4. The genes present in group2 represent conservation in their gene structure. Two genes Gr5G164600, Gr8G16600 are closely related to each other as their intron exon positions are almost similar. In group2 two Arabidopsis genes, At5G50570, At5G50670 are closely related to each other as they have similar position of introns. Moreover, all the genes present in group2 carry 3 introns except one gene which is Gr2G181700. There are 3 genes present in gorup3, two of them At1G76580 and Gr2G112200 have eleven introns while the third gene At1G20980 have nine introns. Among the six genes in group4, two genes Gr4G047200 and Gr5G071400 carry 10 intron while the other 4 genes carry nine intron. Further, the group5 consists of six genes and out of these, 3 carry 9 introns while the other three genes Gr11G029400, At1G53160, At2G33810 carry only one intron. Similarly, one gene (Gr1G176000) from the group number 6 contained three introns while the other two genes Gr8G107400 and At1902065 carry two introns each. In group7, the genes Gr7G087900 and Gr8G107300 carry two introns while the other three genes Gr3G003200, Gr2G068300, Gr13G169800 carry only a single intron. In group 8, At3G15270 contain only one intron while all other five genes carry two intron.

Protein motif analysis for SBP transcription factors in *Arabidopsis* and diploid cotton: Protein motifs of SBP transcription factor proteins were analyzed by online tool MEME (http://meme.nbcr.net/meme/cgi-bin/meme.cgi) which is publically available. The results showed twenty conserved motifs in SBP proteins in both plant species (Fig. 5). The regular expression of highly conversed motif (Motif 1) is PRCQV[ED] GCN [AV]DL[ST] [NS] AKDYHRRH [KR]VCEVH[SA]K[AT][PS]K[VA][IL][VI][AGN]G[LI]EQ RFCQQCSRFHLLSEFDEGKRSCR[RK]RLAGHN[ER]RR RKPQP[DE] which was present in all the protein sequences except Gr5G146800 from Gossypium.

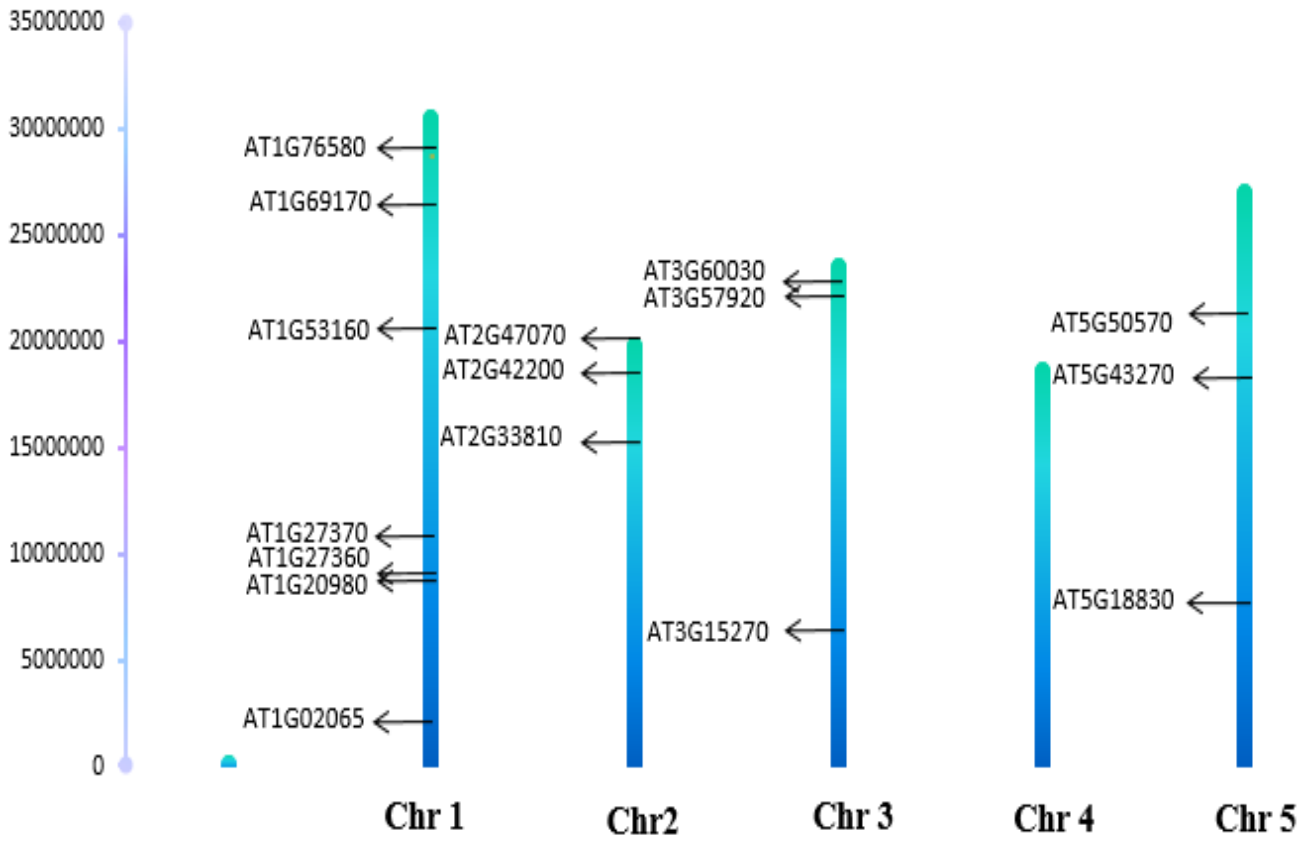

Fig. 2. Chromosomal mapping; of *Arabidopsis thaliana SBP* gene family. Scale illustrated the positions of genes. It contains five chromosomes. Arrow head indicates the positions of genes on each chromosome.

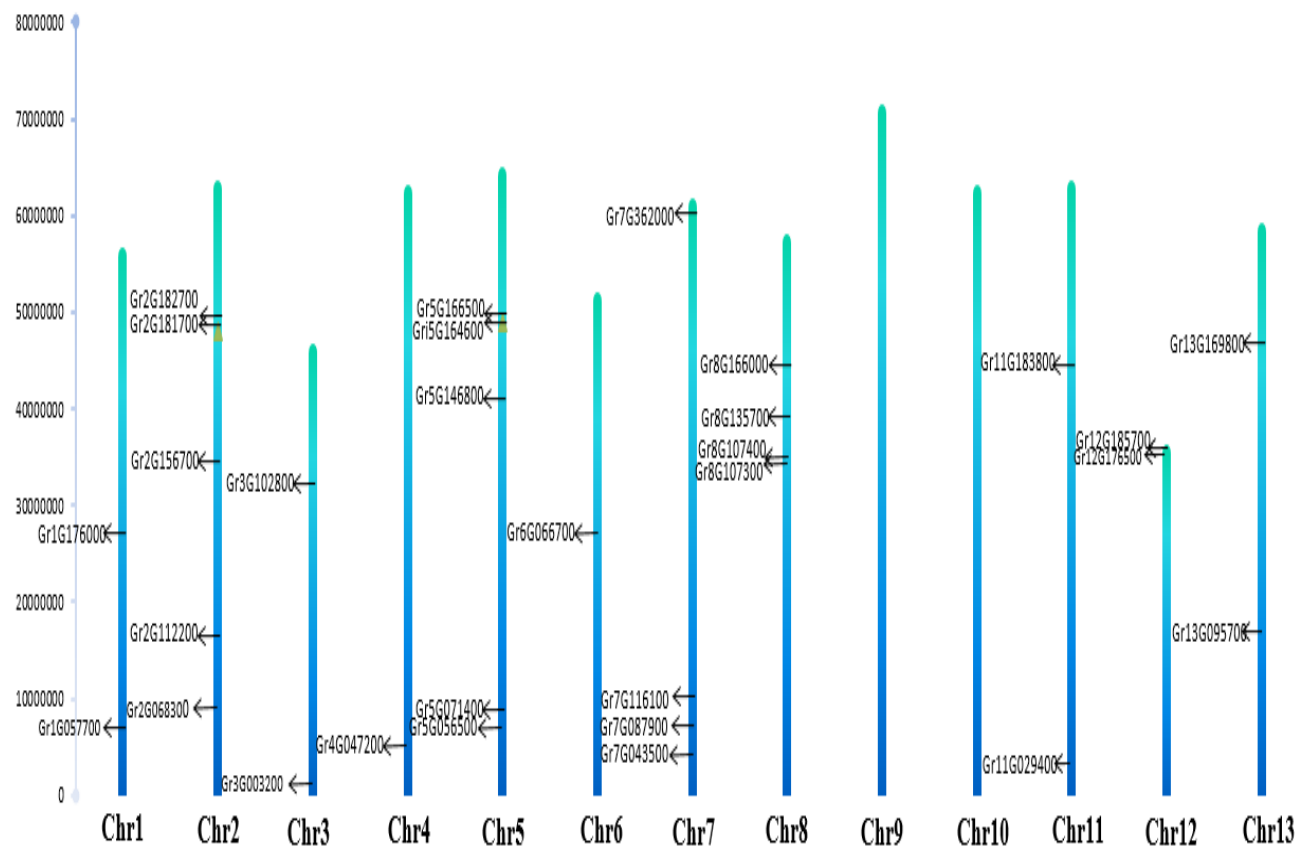

Fig. 3. Chromosomal mapping of *Gossypium raimondii SBP* gene family. Scale illustrated the positions of genes. It contains 13 chromosomes. Arrow head indicates the positions of genes on each chromosome.

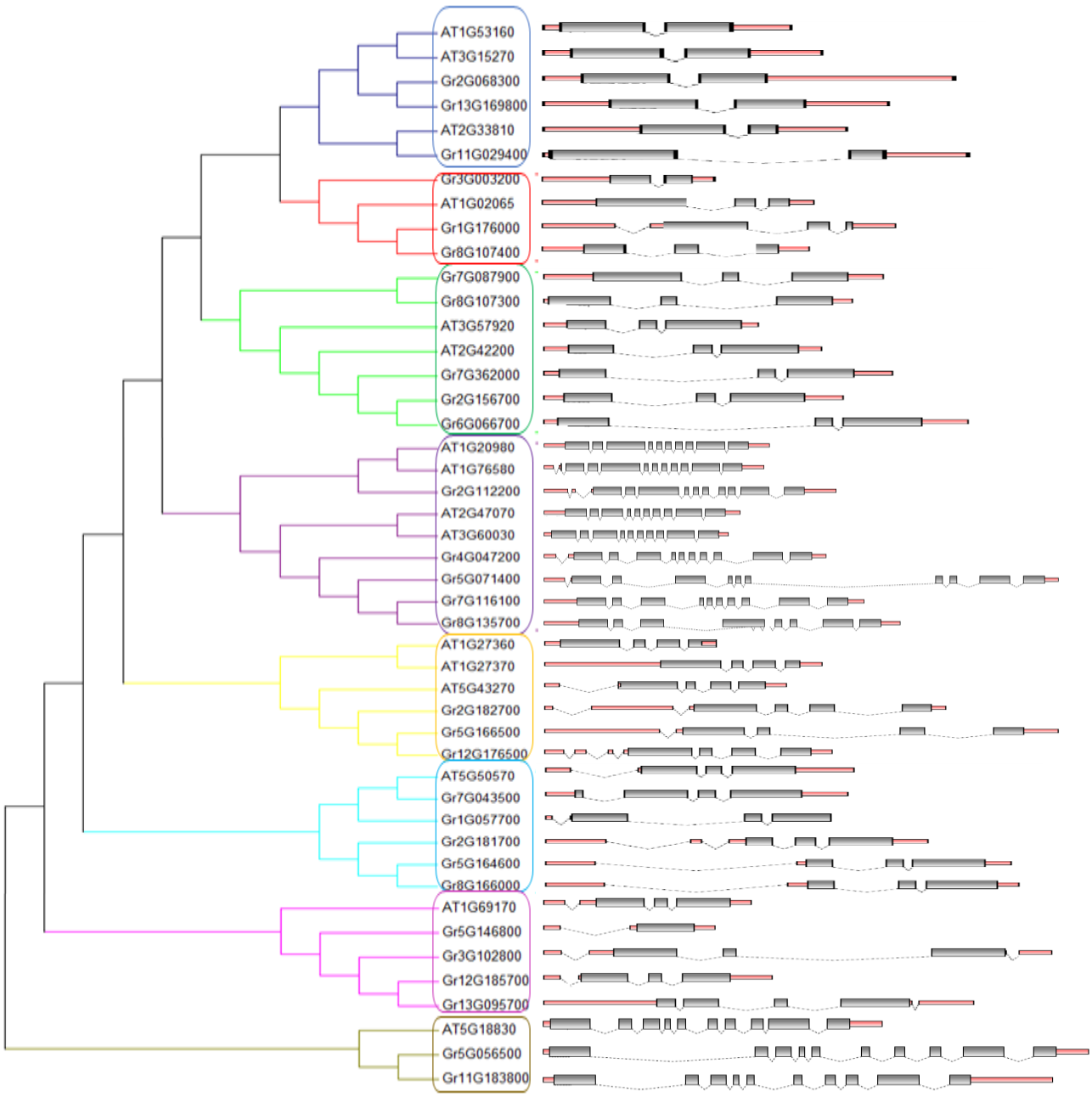

Fig. 4. Phylogenetic relationship and gene structure analysis of *Arabidopsis thaliana* and *Gossypium raimondii SBP* genes. Pink and grey boxes represent UTR and exons respectively and introns are represented by black doted lines. The size of UTR's, exons and introns can be estimated by scale. The analysis was performed by fancy gene.

The SBP proteins present in group1 represent higher conservation of protein sequences as all of them contain motif 1, except one protein Gr5G146800 and motif 9 except Gr5G146800. However, motif 1 and motif11 is present in only one gene of group1 which is Gr5G146800. Further, motif15 is present in 3 proteins Gr5G166500, Gr12G176500 and Gr2G182700. The genes present in group2 are closely related to each other as all of them contain motif 1 and motif 9. The protein sequences in the group3 carry motif 1, 4, 5, 6, 8, 10, 14, 16, 17and18. The presence of these motifs shows these protein sequences are highly conserved and have lot of information to be transacted. Two proteins of group2 At1G20980 and Gr2G112200 also carry motif 9. The protein sequences in group4 are also conserved within the cluster as motif 1,6,7,8,10,12,13,14,16,17 and 18 are present in all sequences except one sequence (At3G60030) which does not contain motif 16. The sequences present in group5 do not show considerable conservation. Three proteins from this group Gr5G056500, Gr11G183800, At5G18830 carry motif 1, 2, 5, 14 and 16, while the other three genes Gr11G029400, At1G53160, At2G33810 carry motif 1only. The protein sequences of group 6 are short sequences and carry two motifs i.e. motif 1 and motif 11. The two proteins sequences of group 7 carry motif 1, 9 and 11 along these motifs Gr7G087900 also carry motif 14 and Gr8G107300 also carry motif 8 as shown in Fig. 5. The other three protein sequences Gr3G003200, Gr2G068300 and Gr13G169800 carry only motif 1. All the protein sequences of group 8 carry motif 1, 9 and 11 except one protein sequence At3G15270 which consist of only one conserved region, the motif number 1.

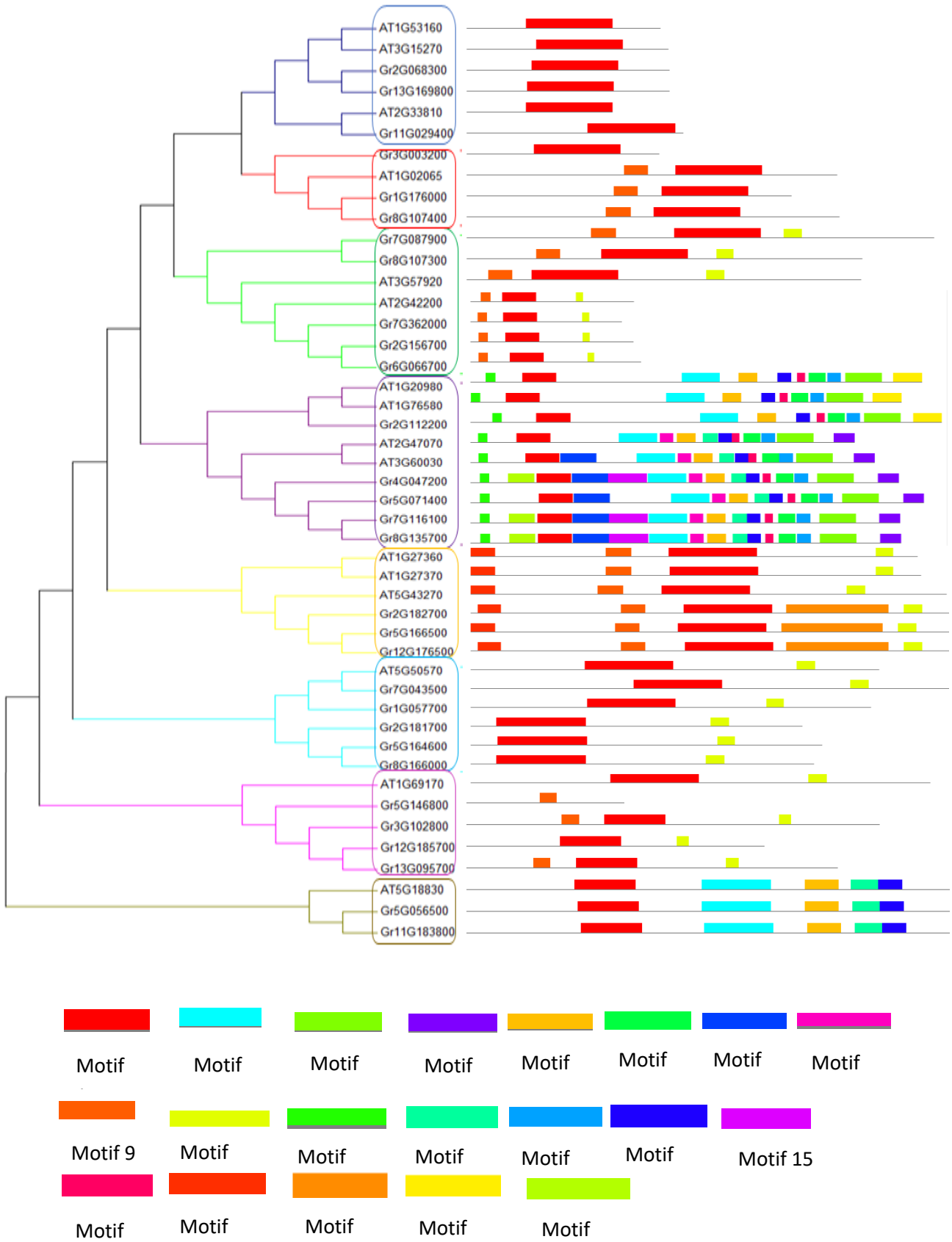

Fig. 5. Phylogenetic relationship and motif analysis of *Arabidopsis thaliana* and *Gossypium raimondii* genes. A multiple alignments of full length amino acids from Arabidopsis and Gossypium were accomplished by Clustal W and the phylogenetic tree was established using MEGA 6 by the neighbor joining (NJ) method with 300 bootstrap replicates. Simplified description of the conserved motifs in the *SBP* genes family from *Arabidopsis thaliana* and *Gossypium raimondii* were clarified by MEME. The colored box in each line represent motif. The non-conserved motifs are represented by blank lines.

Multiple sequence alignment of SBP domains of *Arabidopsis* and *Gossypium raimondii*: The signature domain from SBP proteins in both plant species was aligned to see the similarities, differences and conservation of various amino acid residues using multiple sequence alignment. The length of the domain was assessed from plant transcription factors database (TFDB) site versio3.0. (http://planttfdb.cbi.pku.edu.cn/). Multiple sequence alignment of Arabidopsis Thaliana domain is shown in (Figs. 6 and 7). In Arabidopsis domain, 24 residues out of 76 residues of SBP domain were highly conserved while in Gossypium 10 residues out of 76 residues were highly conserved.

Analysis of *Cis* regulatory elements in the promoters of SBP genes from both plant species: The expression level of genes is regulated by cis regulatory elements present in the promoter region. The *cis* elements were determined from 1000bp upstream to the start codon in the promoter region of SBP transcription factors. The presence of five *cis* elements i.e. TATAbox, GATAbox, Ebox, Wbox, CCAATbox and CAATbox was evaluated in this region of the promoters. The number and location of different elements is given in Table 2, Figs. 8 and 9. In Arabidopsis, the promoter of most of the SBP genes contained higher number of these *cis* elements between 300bp to 780bp (Fig. 8). The promoter sequence of all the genes carry GATAbox and Ebox. Similarly, the promoters of all the SBP genes from *Arabidopsis* carry Wbox except two genes which are At1G53160 and At5G18830. It was also perceived that there was one gene At2G33810 the promoter of which does not include TATAbox. The promoter of At3G60030 does not contain CAATbox, while those of At3G57920 and At3G60030 do not include CCAATbox.

However, in *Gossypium raimondii* CAATbox and TATAbox are present in in the 1kb promoter sequence all the genes except Gr3G071400 which does not have TATAbox in it (Fig. 9). Ebox is also present in all SBP promoters except the promoters of Gr7G087900 and Gr7G116100 genes. The promoter region of 13 genes in Gossypiun do not contain CCAATbox which are Gr2G112200, Gr2G181700, Gr3G071400, Gr7G087900, Gr7G116100, Gr7G362000, Gr8G107300, Gr8G107400, Gr8G135700, Gr8G166000, Gr11G029400, Gr11G183800, Gr12G185700. Similarly, Wbox is not present in four promoters Gr7G087900, Gr8G166000, Gr11G029400 and Gr12G176500. GATAbox is absent in the promoter sequences of two genes Gr12G185700 and Gr13G095700.

Evolutionary relationship of *Arabidopsis* and *Gossypium raimondii* SBP genes: To get the idea about the origin and evolutionary relationship of the SBP protein family genes in these plants, a comparative synteny analysis between Arabidopsis and Gossypium SBP protein sequences was performed. This synteny analysis was performed between 17 Arabidopsis and 30 Gossypium SBP proteins. This synteny analysis represents that the proteins from both species are closely related to each other and they show higher similarity but there are some genes which have greater similarty than the other genes (Fig. 10). This analysis represent that a single Gossypium gene Gr2G112200 is syntenic to multiple Arabidopsis genes such as -At1G76580 and At1G20980. Similarly, there are some genes of *Arabidopsis thaliana* which corresponds to multiple *Gossypium raimondii* genes such as At2G47070-Gr5G071400/ Gr7G116100/ Gr8G135700, At3G60030-Gr5G071400/Gr7G116100/ Gr8G135700/Gr4G047200 as shown in (Fig. 10).

Discussion

It has been reported that SBP box genes are only found in plants (Cardon *et al*., 1999; Guo *et al*., 2008). The SBP genes present in green plants ranging from unicellular green algae, Lycophytes and mosses to angiosperms and gymnosperm. While golden algae, red algae and brown algae do not carry SBP-box genes (Guo *et al*., 2008). SBP transcription factor gene family has been the subject of intensive studies, as they have been identified in Rice (Xie *et al*., 2006; Yang *et al*., 2008), Tomatto, Maize and Arabidopsis (Cardon *et al*., 1999; Yang *et al*., 2008), *Physcomitrella patens* (Riese *et al*., 2007), Grapevine (Hou *et al*., 2010; Wang *et al*., 2010), *Malus domestica* (Li *et al*., 2013), Castor Bean (Zhang & Ling, 2014) and *Gossypium hirsutum* (Zhang *et al*., 2014b).

In this study we performed a comparative computational analysis of SBP proteins in the model plant Arabidopsis and diploid cotton (*Gossypium raimondii*). The phylogenetic tree revealed 8 groups of SBP proteins on the basis of protein sequence, intron-exon structure and protein motif analysis. Guo *et al*. (2008) carried out a phylogenetic analysis of SBP genes from green algae, mosses, platens, gymnosperms, dicotyledoneous angiosperm Arabidopsis and the monocotyledoneous angiosperm rice (*Oryza sativa*) and maize (Zea mays).The results represented that all the sequences from the green algae were grouped separately in one clade while all the other sequences from the land plants were present in other group which represents the similarity between land plants (Guo *et al*., 2008). Furthermore, Li *et al*. 2013 also carried out a phylogenetic relationship by using SBP-box genes from green algae, *P. paten*, moss, apple, grapes, tomato, Arabidopsis and the results showed that all the sequences from land plants were present in one group while the sequences from the green algae were clustered into other clad. Wang *et al*. (2009) carried out a phylogenetic analysis by using the SBP-box genes from grapevine and Arabidopsis and the resulted tree was divided into three groups A,B,C. The group A was further subdivided into six subgroups and the group B was subdivided into two groups in order to identify the orthologues and paralogues relationship. As the sequences from both grapevine and Arabidopsis have similarity so it was concluded that they may be originated from the same ancestor plants before the divergence of grapevine and Arabidopsis. In our study a phylogenetic tree was generated based on the SBP proteins identified in two land plant Arabidopsis and diploid cotton, the tree was divided into eight groups and the sequences present in each group represents similarity between sequences. From phylogenetic analysis, it was observed that the genes from Arabidopsis and cotton were not clustered in a species specific manner as the genes from both plants were scattered throughout the phylogenetic tree. In addition, it was also observed that each group contains SBP protein sequences from both plants which suggest that the SBP protein sequences from these plants have certain extent of similarity between them.

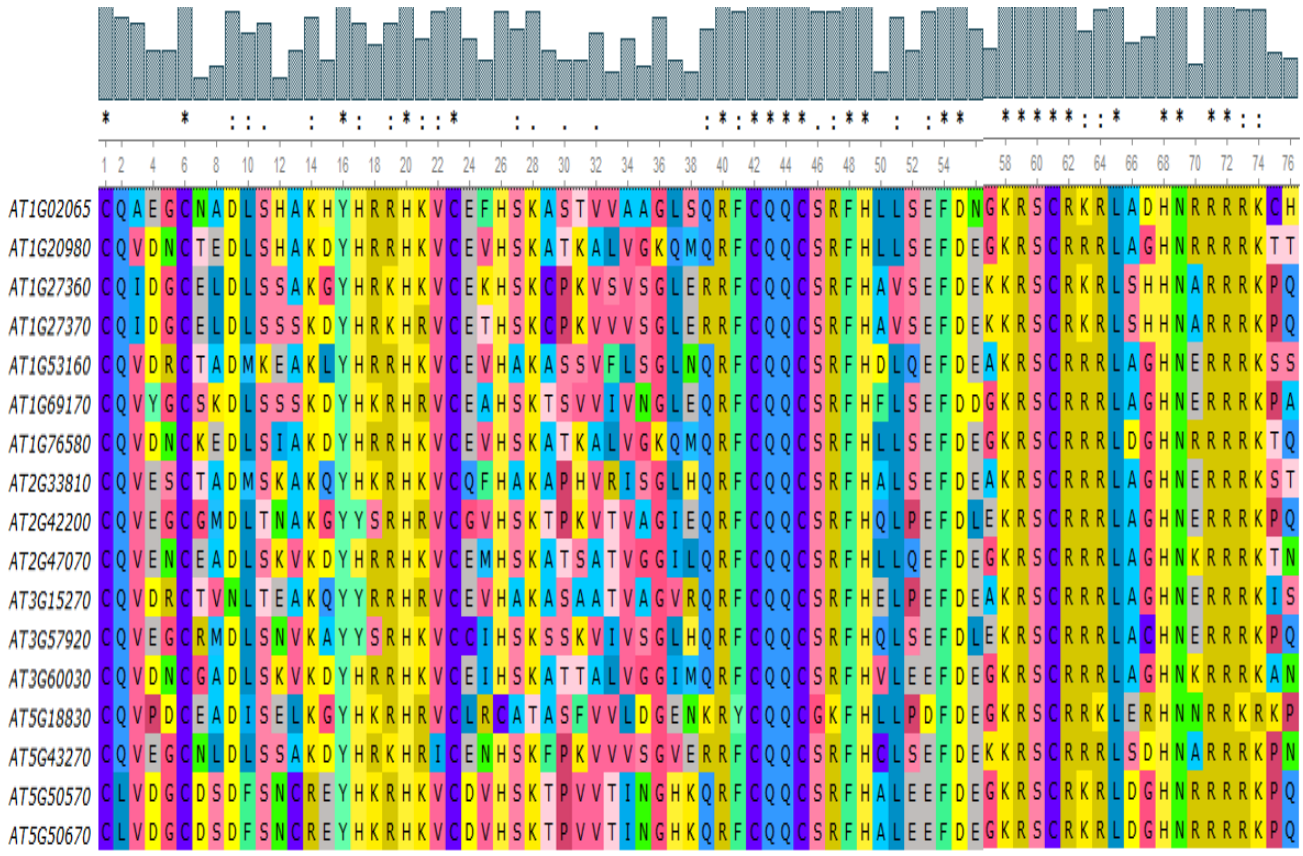

Fig. 6. Multiple sequence alignment of *Arabidopsis thaliana* SBP domain. The alignment was carried out by using unipro UGENE. Alignment contains 26 conserved residues.

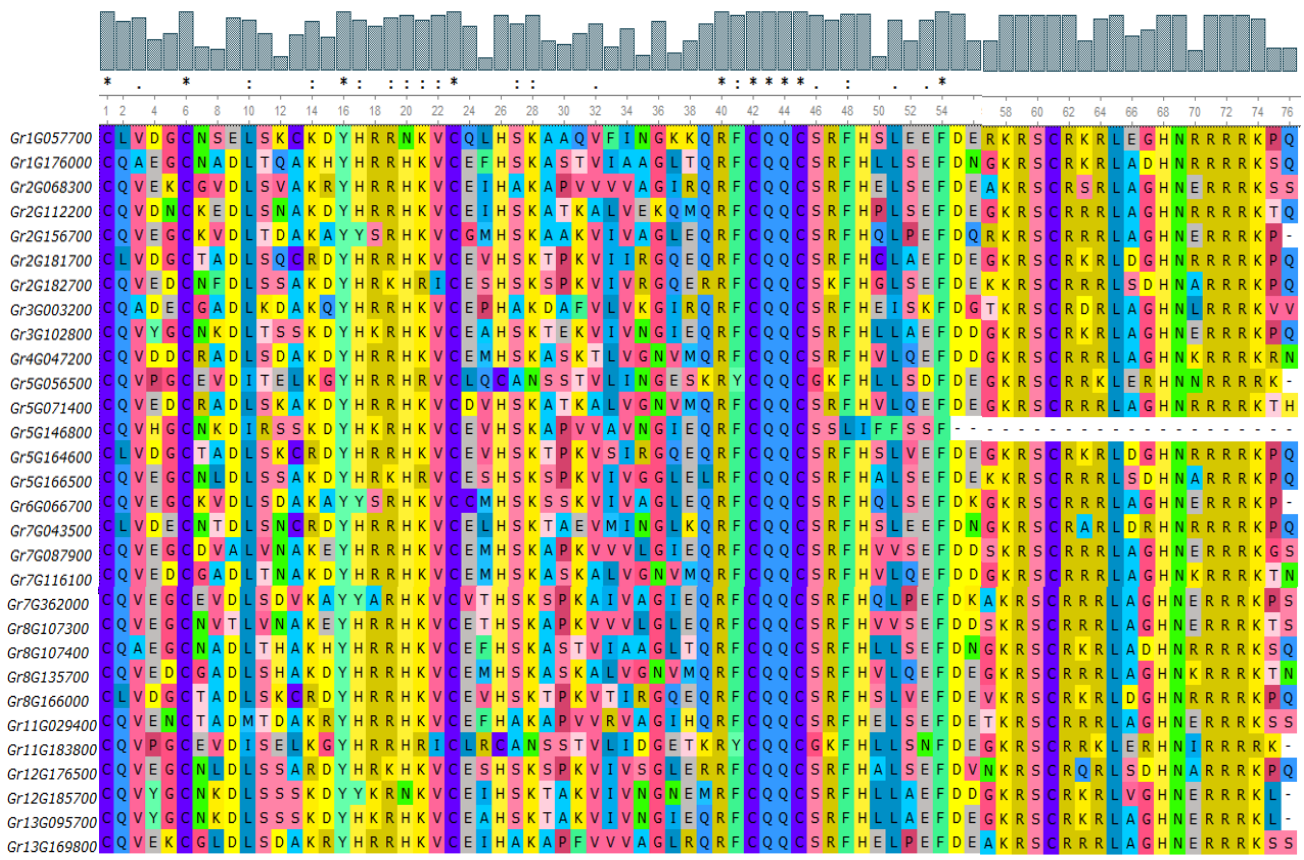

Fig. 7. Multiple sequence alignment of *Gossypium raimondii* SBP domain. Alignment was carried out by using unipro UGENE. Alignment contains 16 conserved residues.

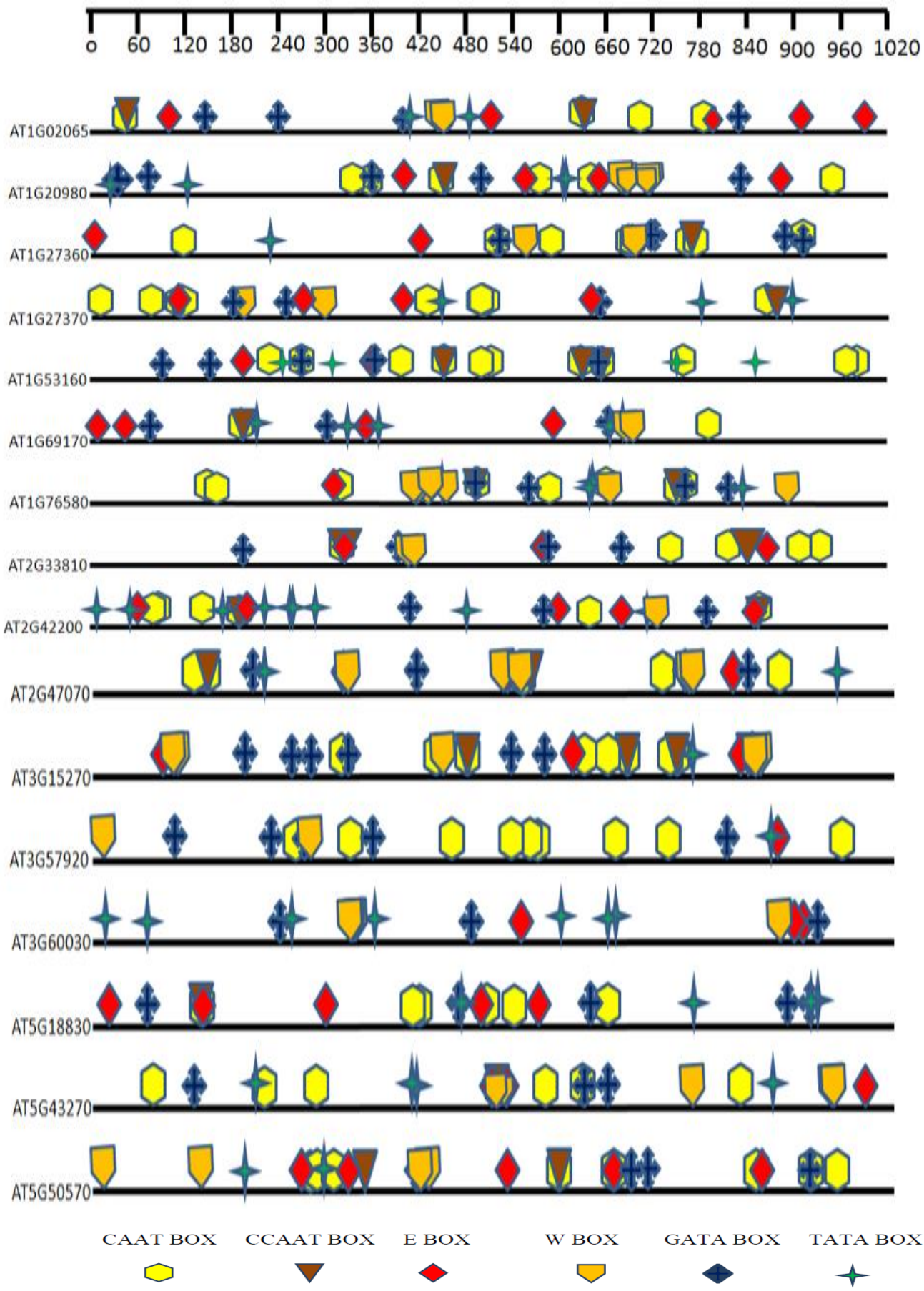

Fig. 8. Promoter analysis of *Arabidopsis thaliana SBP* gene family. The analysis was performed by PLACE, a database of nucleotide sequence motifs found in Plant Cis-acting regulatory DNA Elements. Binding sites are represented by following symbols.

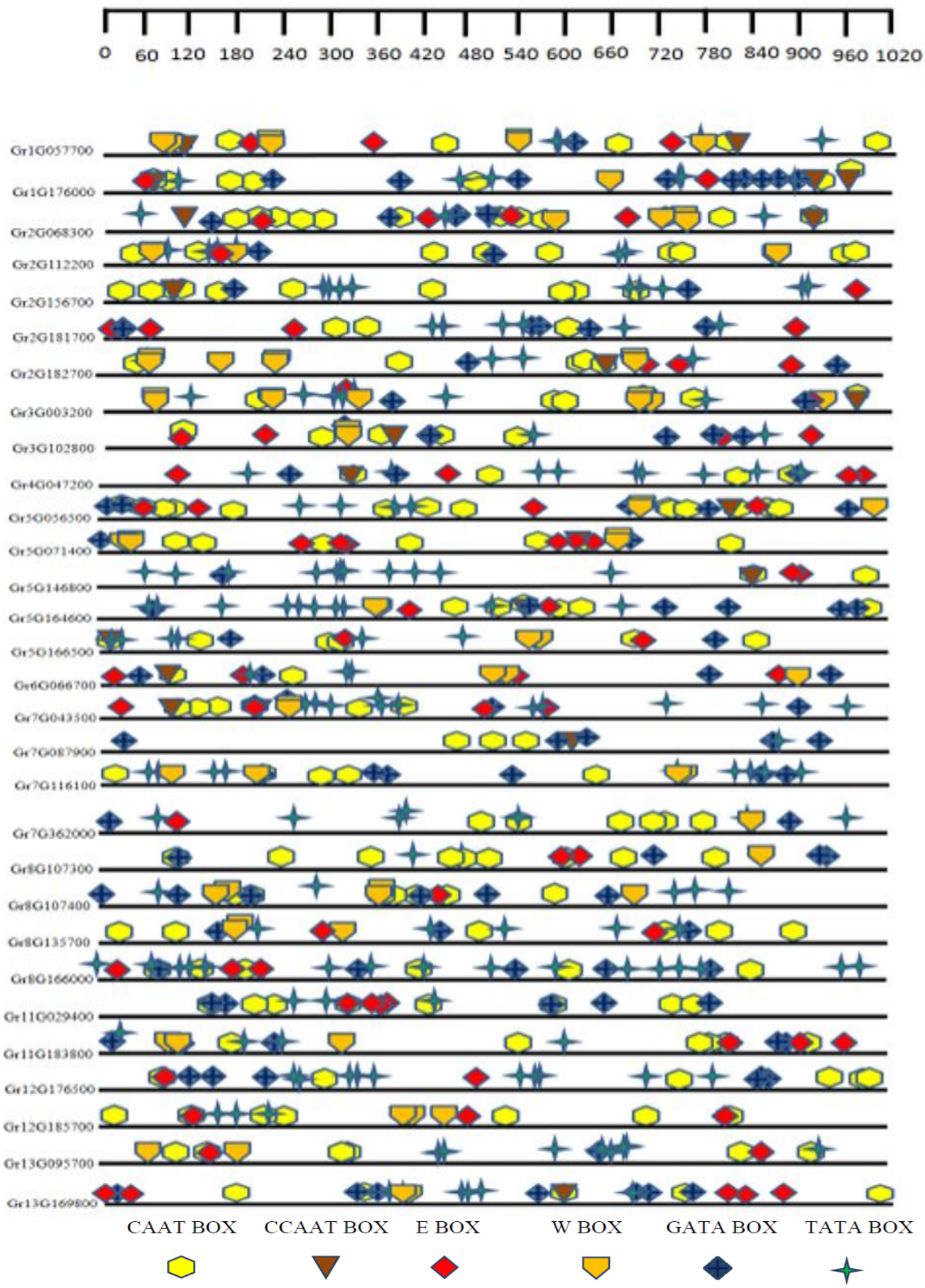

Fig. 9- Promoter analysis of *Gossypium raimondii* SBP gene family. The analysis was performed by PLACE, a database of nucleotide sequence motifs found in Plant *Cis*-acting regulatory DNA Elements. Binding sites are represented by following symbols.

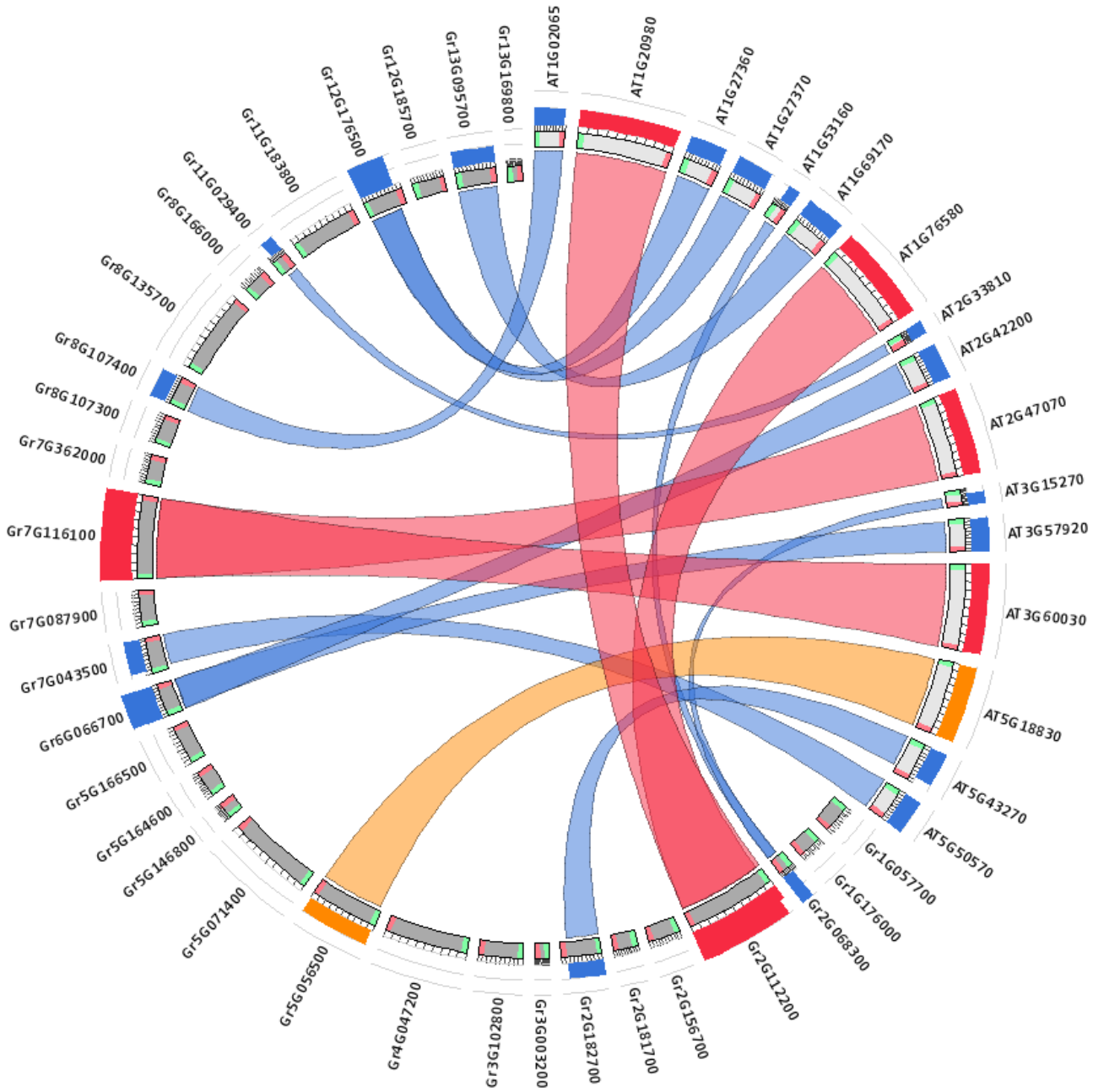

Fig. 10. Synteny analysis of SBP-box genes between *Arabidopsis* and Gossypium. Colored lines which connect two regions indicate syntenic regions between *Arabidopsis* and *Gossypium*.

The genomic sequences of SBP genes present in rice contain long introns. However, the Arabidopsis SBP genes have average length of intron 124bp and Rice have average length of 520bp (Guo *et al*., 2008). Guo and his colleagues carried out the gene structure analysis by using SBP-box genes of Arabidopsis and rice and the results shows that the genes present in the same group have almost similar gene structure with a little exception due to variation in number of exons and introns. Furthermore, Li *et al*. (2013) also carried out a gene structure analysis of SBP factors from apple and proposed that the sequences present in same group almost have similar exon-intron structure, similar number of exons and introns with little exception. In our study, the gene structure analysis was carried out for the SBP-box genes from Arabidopsis and cotton. The results reveal that the sequences present in one group carries almost similar gene structure with some exceptions. Moreover, those SBP genes which have similar exon-intron pattern might have originated from a common ancestor. Whereas, the difference in exon-intron structure among different groups hints that these SBP factors may have different physiological functions. As in early studies it was reported that SBP box genes from different groups play different roles I, in flower and fruit development, in response to copper and fungal toxin and in sporogenesis (Guo *et al*., 2008).

This study also analyzes conserved protein motifs of SBP transcription factors which demonstrated conservation of several motifs in different groups. The protein sequence of the genes present in same group have almost similar pattern of conserved motifs. The motif 1 is present in all Arabidopsis and cotton SBP proteins as it

carries the SBP signature domain sequence which is two zinc binding sites and a nuclear localization signals sequence. Along with motif 1 there are other motifs which are conserved in different groups but the function of these motifs is not known, however, they might be considered as structural units (Zhang *et al*., 2014a). In SBP domain analysis we identified two zinc finger like structures or zinc binding sites one zinc binding site (Zn1) contain three cysteine and one histidine residue (Cx4Cx16Cx2H) while second zinc binding site (Zn2) contains three cysteine and one histidine residue (Cx2Cx3Hx11C) along these two zinc binding site the SBP domain also carry a highly conserved region known as nuclear localization signal (NLS) in overlapping with Zn2. The consensus sequence for NLS is KRx11RRRK. In previous studies these two zinc finger like structures and NLS was also observed as conserved part of the SBP domain in Grapevine (Wang *et al*., 2010) and in Arabidopsis (Zhang *et al*., 2014a). The conservation of zinc binding sites and the NLS in SBP domain is considered important for specific recognition and binding to *cis*-elements in the promoter of target genes (Zhang *et al*., 2014a).

The physical distribution of the SBP genes on the chromosomes in various plants has also been reported previously. For example, in apple all the 27 genes were distributed among nine out of the seventeen chromosomes which were chromosome 3, 6, 7, 9, 11, 13, 14, 16, 17 with maximum number of 8 genes on chromosome 13 and 16 (Li *et al*., 2013). In our study in *Arabidopsis*, 16 SBP genes were distributed among all the five chromosomes with a maximum of 7 genes on chromosome 1 and in *Gossypium raimondii* all the thirty genes are distributed among thirteen chromosomes with the maximum number of 5 SBP genes on chromosome 2 and 5 each.

Six *cis*- acting regulatory elements, CAAT-BOX, CCAAT-BOX, E-BOX, W-BOX, GATA-BOX and TATA-BOX, were located and mapped on 1 kb upstream of initiation codon. They all are involved in gene regulation under stress conditions. The consensus signature sequences of these *cis* elements are CAAT, GGCCAATCT, CACGTG, TGAC, TATAAA and GATA respectively. The extent of the specificity of gene expression depends on *cis*-regulatory elements and their binding and interaction with the transcription factor (Zhou, 1999). The GT1-BOX interacts with GATA element and gives light-induced expression, however, its binding with other light responsive elements is necessary for expression (Zhou, 1999). W-BOX is DNA binding place for WRKY transcription factors which are involved in gene regulation of many processes like plant growth and development, leaf senescence, cell signaling and in response to several biotic and abiotic stress responses (Ali *et al*., 2014; Chi *et al*., 2013; Rushton *et al*., 2012; Rushton *et al*., 2010). Similarly, TATA box is considered very important for transcriptional activation and regulation for variety of processes in plants (Bernard *et al*., 2010). In this study, the analysis of *cis* regulatory element revealed that in Arabidopsis, the promoter sequences of SBP gene have higher number of cis regulatory elements between 300bp to 780bp and all the genes carry GATAbox and Ebox. While in *Gossypium raimondii* CAAT box is present in all genes**.** This indicated that SBP transcription factors are regulated by a variety of other transcription factors. Most of these *cis* regulatory elements are related to transcription factors involved in diverse plant functions which suggest that the regulation of SBP factors is important phenomenon.

Our study also reveals that several Arabidopsis SBP genes were syntenic to those of diploid cotton genes which demonstrated an evolutionary relationship between these genes. According to the synteny analysis between Apple and Arabidopsis shows, 11 apple and Arabidopsis genes were located in the duplicated genomic regions of both species which indicates that many of apple and Arabidopsis genes are derived from common ancestor (Li *et al*., 2013). It has been observed that duplications, including segmental duplication, tandem duplication and genomic duplication play important role in the evolution (Li *et al*., 2013).

These findings will help in identifying and understanding the evolutionary relationships among the SBP transcription factors in different plant species. The bioinformatics analysis of the SBP transcription factor family conducted in the present study provides an overall picture of the classification of SBP family members in Arabidopsis and *Gossypium raimondii*.

References

Ali, M.A., K. Wieczorek, D.P. Kreil and H. Bohlmann. 2014. The beet cyst nematode Heterodera schachtii modulates the expression of WRKY transcription factors in syncytia to favour its development in Arabidopsis roots. *Plos One*. 9:e102360.

Bernard, V., V. Brunaud and A. Lecharny. 2010. TC-motifs at the TATA-box expected position in plant genes: a novel class of motifs involved in the transcription regulation. *BMC Genomics*., 11:166.

Birkenbihl, R.P., G. Jach, H. Saedler and P. Huijser. 2005. Functional dissection of the plant-specific SBP-domain: overlap of the DNA-binding and nuclear localization domains. *J. Mol. Biol*., 352: 585-596.

Cardon, G., S. Hohmann, J. Klein, K. Nettesheim, H. Saedler and P. Huijser. 1999. Molecular characterisation of the Arabidopsis SBP-box genes. *Gene*., 237: 91-104.

Cardon, G.H., S. Hohmann, K. Nettesheim, H. Saedler and P. Huijser. 1997. Functional analysis of the Arabidopsis thaliana SBP-box gene SPL3: A novel gene involved in the floral transition. *Plant J*., 12: 367-377.

Chi, Y., Y. Yang, Y. Zhou, J. Zhou, B. Fan, J.-Q. Yu and Z. Chen. 2013. Protein–protein interactions in the regulation of WRKY transcription factors. *Mol. Plant*., 6: 287-300.

Guo, A.Y., Q.H. Zhu, X. Gu, S. Ge, J. Yang and J. Luo. 2008. Genome-wide identification and evolutionary analysis of the plant specific SBP-box transcription factor family. *Gene*., 418: 1-8.

Hou, H., J. Li, M. Gao, S.D. Singer, H. Wang, L. Mao, Z. Fei and X. Wang. 2010. Genomic organization, phylogenetic comparison and differential expression of the SBP-Box family genes in grape. *PloS One*. 8: e59358.

Imran, M. and J.Y. Liu. 2016. Genome-wide identification and expression analysis of the malate dehydrogenase gene family in *Gossypium arboreum*. *Pak. J. Bot*., 48: 1081-1090.

Khan, A.M., A.A. Khan, M.T. Azhar, L. Amrao and H.M.N. Cheema. 2016. Comparative analysis of resistance gene analogues encoding NBS-LRR domains in cotton. *J. Sci. Food Agri.*, 96: 530-538.

Klein, J., H. Saedler and P. Huijser. 1996. A new family of DNA binding proteins includes putative transcriptional regulators of theAntirrhinum majus floral meristem identity gene SQUAMOSA. *Mol. Gener. Genet.*, 250: 7-16.

Lannenpaa, M., I. Janonen, M.H. Vuori, M. Gardemeister, I. Porali and T. Sopanen. 2004. A new SBP-box gene *BpSPL1* in silver birch (*Betula pendula*). *Physiol Plant*, 120: 491-500

Li, J., H. Hou, X. Li, J. Xiang, X. Yin, H. Gao, Y. Zheng, C.L. Bassett and X. Wang. 2013. Genome-wide identification and analysis of the SBP-box family genes in apple *Malus domestica* Borkh.. *Plant Physiol. Biochem*., 70: 100-114.

Noguero, M., R.M. Atif, S. Ochatt and R.D. Thompson. 2013. The role of the DNA-binding One Zinc Finger (DOF) transcription factor family in plants. *Plant Sci.*, 209: 32-45.

Preston, J.C. and L.C. Hileman. 2013. Functional evolution in the plant squamosa-promoter binding protein-like SPL gene family. *Frontiers Plant Sci*., 4: 80.

Riechmann, J.L. and O.J. Ratcliffe. 2000. A genomic perspective on plant transcription factors. *Curr. Opinion in Plant Biol.*, 3: 423-434.

Riese, M., S. Hohmann, H. Saedler, T. Munster and P. Huijser. 2007. Comparative analysis of the SBP-box gene families in *P. patens* and seed plants. *Gene*., 401: 28-37.

Rushton, D.L., P. Tripathi, R.C. Rabara, J. Lin, P. Ringler, A.K. Boken, T.J. Langum, L. Smidt, D.D. Boomsma, N.J. Emme, X. Chen, J.J. Finer, Q.J. Shen and P.J. Rushton. 2012. WRKY transcription factors: key components in abscisic acid signalling. *Plant Biotechnol. J*., 10: 2-11.

Rushton, P.J., I.E. Somssich, P. Ringler and Q.J. Shen. 2010. WRKY transcription factors. *Trends Plant Sci*., 15: 247-258.

Wang, Y., Z. Hu, Y. Yang, X. Chen and G. Chen. 2009. Function annotation of an SBP-box gene in *Arabidopsis* based on analysis of co-expression networks and promoters. *Int. J. Mol. Sci.*, 10: 116-132.

Wang, Y., Z. Hu, Y. Yang, X. Chen and G. Chen. 2010. Genome-wide identification, phylogeny, and expression analysis of the SBP-box gene family in grapevine. *Russ. J. Plant Physiol*., 57: 273-282.

Watanabe, K.N. and M.A. Khan. 2016. Bio-informatic analysis of a vacuolar Na^+ /H^+ antiporter (*ALaNHX*) from the salt resistant grass *Aeluropus lagopoides*. *Pak. J. Bot*., 48: 57-65.

Xie, K., C. Wu and L. Xiong. 2006. Genomic organization, differential expression, and interaction of SQUAMOSA promoter-binding-like transcription factors and microRNA156 in rice. *Plant Physiol*., 142:280-293.

Yamasaki, K., T. Kigawa, M. Inoue, M. Tateno, T. Yamasaki, T. Yabuki, M. Aoki, E. Seki, T. Matsuda, E. Nunokawa, Y. Ishizuka, T. Terada, M. Shirouzu, T. Osanai, A. Tanaka, M. Seki, K. Shinozaki and S. Yokoyama. 2004. A novel zinc-binding motif revealed by solution structures of DNA-binding domains of Arabidopsis SBP-family transcription factors. *J. Mol. Biol*., 337: 49-63.

Yang, Z., X. Wang, S. Gu, Z. Hu, H. Xu and C. Xu. 2008. Comparative study of SBP-box gene family in *Arabidopsis* and rice. *Gene*. 407:1-11.

Zhang, L., B. Wu, D. Zhao, C. Li, F. Shao and S. Lu. 2014a. Genome-wide analysis and molecular dissection of the SPL gene family in Salvia miltiorrhiza. *J. Integr. Plant Biol*., 56: 38-50.

Zhang, S.D. and L.Z. Ling. 2014. Genome-wide identification and evolutionary analysis of the SBP-Box Gene family in castor bean. *PloS one*. 9:e86688.

Zhang, X., L. Dou, C. Pang, M. Song, H. Wei, S. Fan, C. Wang, and S. Yu. 2014b. Genomic organization, differential expression, and functional analysis of the SPL gene family in Gossypium hirsutum. *Mol. Genet. Genom*., 2901: 115-26.

Zhou, D.X. 1999. Regulatory mechanism of plant gene transcription by GT-elements and GT-factors. *Trends Plant Sci*., 4: 210-214.

(Received for publication 13 April 2016)