Available online at www.sciencedirect.com
ScienceDirect
Gene xx (2007) xxx – xxx
GENE
SECTION
EVOLUTIONARY GENOMICS
www.elsevier.com/locate/gene

Comparative study of SBP-box gene family in *Arabidopsis* and rice

Zefeng Yang [1], Xuefeng Wang [1], Shiliang Gu, Zhiqiu Hu, Hua Xu, Chenwu Xu *

Jiangsu Provincial Key Laboratory of Crop Genetics and Physiology, Key Laboratory of Plant Functional Genomics of Ministry of Education, Yangzhou University, Yangzhou, 225009, China

Received 6 November 2006; received in revised form 5 February 2007; accepted 8 February 2007

Received by I.B. Rogozin

Abstract

SBP-box proteins are plant-specific putative transcription factors, which contain highly conserved SBP domain and could bind specifically to promoters of the floral meristem identity gene *SQUAMOSA* and its orthologous genes to regulate their expressions. In this study, 17 non-redundant SBP-box genes in *Arabidopsis* genome and 19 in rice genome were identified by using the known SBP domain sequences as queries. The phylogenetic analysis suggested that the main characteristics of this family might have been in existence before the split of *Arabidopsis* and rice, and most SBP-box genes expanded in a species-specific manner after the split of monocotyledon and dicotyledon. All the SBP-box proteins were classified into 9 subgroups based on the phylogenetic tree, where each group shared similar motifs and the orders of the motifs in the same group were found almost identical. Analysis of nonsynonymous and synonymous substitution rates revealed that the SBP domain had gone through purifying selection, whereas some regions outside SBP domain had gone through positive or relaxed purifying selection. The expression patterns of the SBP-box genes were further investigated by searching against the EST database. Results showed that the *Arabidopsis* SBP-box genes are expressed chiefly in flowers, leaves, roots and seeds, while those in rice mainly in flowers and callus.

Keywords: Transcription factor; Phylogenetic tree; EST; Nonsynonymous and synonymous substitution rate

1. Introduction

Transcription factors (TFs) are DNA-binding proteins that regulate gene expression at the level of mRNA transcription. They are capable of activating or repressing transcription of multiple target genes (Riechmann et al., 2000; Xiong et al., 2005). SBP-box gene family belongs to the type of plant-specific zinc finger protein genes, which encode putative plant-specific transcription factor. SBP-box genes were first identified in *Antirrhinum majus* for the capacity of their protein products to bind to promoter region of the floral meristem identity gene *SQUAMOSA* (Klein et al., 1996). Since then, SBP-box genes had been found in diverse plant species and their functions had been extensively investigated (Becraft et al., 1990; Cardon et al., 1997; Shao et al., 1999; Lannenpaa et al., 2004). Birkenbihl et al. found that there were 16 SBP-box genes in the model species *Arabidopsis thaliana,* named as *SPL1* to *SPL16* (Birkenbihl et al., 2005). Their protein products could bind specifically to the related motifs in the promoter of *AP1* (the orthologous gene in *Arabidopsis* of *SQUAMOSA*). *SPL3* was found highly expressed in vegetative and inflorescence apex, floral meristem, leaf and floral primordial (Cardon et al., 1997). *SPL3*, *SPL4* and *SPL5* showed dramatically up-regulated in response to long day floral induction (Schmid et al., 2003). And *SPL8* showed to affect pollen sac development (Unte et al., 2003). One of the largest SBP-box genes, *SPL14*, was recently characterized as conferring resistance to the programmed cell death (PCD)-inducing fungal toxin fumonisin B1 (FB1) (Stone et al., 2005). Another two SBP-box genes were isolated from maize, known as *LIGULELESS1* (*LG1*) (Becraft et al., 1990) and *TEOSINTE GLUME ARCHITECTURE* (*TGA1*) (Wang et al., 2005a). Becraft et al. reported that in the absence

Abbreviations: cDNA, complementary deoxyribonucleic acid; CDS, coding sequence; Dof, DNA binding with one finger; EST, expressed sequence tag; MEME, multiple EM for motif elicitation; mRNA, messenger ribonucleic acid; NCBI, National Center for Biotechnology Information; RLK, receptor-like kinase; SBP, *squamosa* promoter binding protein; TAIR, the *Arabidopsis* Information Resource; TIGR, the Institute of Genomic Research.

* Corresponding author. Tel.: +86 514 7979358; fax: +86 514 7996817.
E-mail address: qtls@yzu.edu.cn (C. Xu).

[1] These authors contributed equally to this work.

0378-1119/$ - see front matter © 2007 Published by Elsevier B.V.
doi:10.1016/j.gene.2007.02.034

of *LG1* gene expression, ligule and auricle were not formed and a sharp boundary failed to develop between sheath and blade (Becraft et al., 1990). While Wang et al. found that several nucleotide changes in the promoter and coding sequences of *TGA1* were probably responsible for the evolution of maize inflorescence architecture (Wang et al., 2005a).

The SBP-box proteins all contain a highly conserved DNA-binding domain (DBD)-SBP domain of approximately 79 amino acid residues in length. The structural basis for this domain is two Zn-finger like structures (Yamasaki et al., 2004) associated with a highly conserved bipartite nuclear localization signal (NLS) (Birkenbihl et al., 2005). The SBP-box genes are ubiquitous in higher plants and play important roles in plant growth and flower development. Further studies on this family can help not only illustrating the developmental processes of high plants but also elucidating the evolutionary relationships between different species.

Rice is one of the most important food crops and is considered the model organism of monocotyledon for molecular and genetic studies. Compared with *Arabidopsis* and maize, rice SBP-box genes research is still in its babyhood, although there had been some relevant reports (Shao et al., 1999). The draft genome sequences of *Oryza sativa* ssp. *indica* and *O. sativa* ssp. *japonica* were simultaneously released by Beijing Genomics Institute (Yu et al., 2002) and Syngenta Company (Goff et al., 2002) in 2002, and the high quality finished sequences of *O. sativa* ssp. *japonica* were released by the International Rice Genome Sequencing Project (IRGSP) in 2005 (IRGSP, 2005). The completed rice genome provides a new platform for searching unknown genes, investigating their functions and studying developmental and evolutionary biology. In this study, using bioinformatics resources and tools, we compared this plant-specific transcription factor family in dicotyledon and monocotyledon model species, i.e., *Arabidopsis* and rice. A neighbor-joining phylogenetic tree was constructed using full-length protein sequences of the SBP-box genes. Twelve pairs of paralogous SBP-box genes were identified from the tree. To examine the driving force for the gene evolution, nonsynonymous and synonymous substitution rate (*Ka*/*Ks*) analysis for the paralogs was performed. We also searched the Genbank EST database in order to get a better picture of the tissue-specific expression of SBP-box genes in *Arabidopsis* and rice.

2. Materials and methods

2.1. Identification of the members of SBP-box family in Arabidopsis and rice

The presumed genes of SBP-box family in *Arabidopsis* and the encoded protein sequences were obtained from TAIR database (http://www.arabidopsis.org). Pfam software (Sonnhammer et al., 1997) was used to deduce their SBP domains. The predicted sequences were then set as queries to search TIGR *Arabidopsis thaliana* annotation deduced protein database (http://www.tigr.org/tdb/e2k1/ath1/) and the TIGR Rice Annotation database (http://www.tigr.org/tdb/e2k1/osa1/) with the blastp program. If the sequence satisfied $E \leq 10^{-10}$, it was selected a candidate protein. Then the tool of Pfam was used to predict the SBP domains of all these candidate proteins. If there was SBP domain in the candidate protein, it belonged to the SBP-box family. The nucleotide and protein sequences of SBP-box genes were downloaded from the TIGR database after getting their basic information. Sequences for *Oryza sativa* ssp. *indica* were obtained from the whole genome shotgun sequencing project of the Beijing Genomics Institute at NCBI (http://www.ncbi.nlm.nih.gov/BLAST/Genome/PlantBlast.shtml?2).

2.2. Sequence analysis

Multiple sequences alignment of SBP-box proteins was performed using the Clustal X 1.83 program (Thompson et al., 1997) with default parameters after getting the protein sequences of SBP-box genes in *Arabidopsis* and rice. Bootstrapping analysis with a Phylip format tree output was carried out after the neighbor-joining method (Saitou and Nei, 1987) using Clustal X 1.83 program and the phylogenetic tree was represented with the help of the Treeview 1.61 software (Zhai et al., 2002). *Arabidopsis* and rice conserved motifs analysis within the determined SBP-box family was performed by MEME program (http://meme.sdsc.edu) with default settings, except that the maximum number of motifs to find was defined as 40. To predict putative functions of identified motifs, the consensus sequences of these motifs were searched against the PROSITE database using the ScanPROSITE tool (http://expasy.org/tools/scanprosite/).

2.3. Ka and Ks calculations

The paralogs for SBP-box genes in *Arabidopsis* and rice were inferred from the phylogenetic tree. Pairwise alignments of the paralogous nucleotide sequences were performed using the program CLUSTAL X 1.83, with the corresponding protein sequences as the alignment guides. Gaps in the alignments were removed manually. The program K-estimator 6.1 (Comeron, 1999) was used to carry out *Ka* and *Ks* analysis for the paralogous genes.

2.4. Expression analysis based on ESTs

Searches for EST sequences corresponding to the *Arabidopsis* and rice SBP-box genes were made in the Genbank EST database (http://www.ncbi.nlm.nih.gov/dbEST/) using the blastn program and default parameters, and the CDS sequences of SBP-box genes were used as query sequences. The EST sequence with hits rate above 95%, longer than 160 bp and the threshold under 10^{-10} was set as corresponding sequence of the SBP-box gene. The entries identified were checked for tissue origin at the GenBank nucleotide sequence database.

3. Results

3.1. Identification of SBP-box genes in Arabidopsis and rice

3.1.1. SBP-box genes in Arabidopsis

The identification of the members of SBP-box gene family in *Arabidopsis* was performed in three steps. The first step involved

Please cite this article as: Yang, Z. et al. Comparative study of SBP-box gene family in *Arabidopsis* and rice. Gene (2007), doi:10.1016/j.gene.2007.02.034

obtaining the nucleotide sequences and deduced amino acid sequences of 16 known SBP-box genes in *Arabidopsis* from the TAIR database, which were found by Birkenbihl and other research groups (Cardon et al., 1999; Birkenbihl et al., 2005). The second step aimed at a complete search for putative SBP-box proteins in *Arabidopsis* and was performed by blastp searches for the *Arabidopsis* genome annotation database in TIGR with the SBP domain sequences inferred by Pfam as queries. As the last step, each predicted SBP-box protein sequence was confirmed by a Pfam search for conserved SBP domain. Altogether 27 non-redundant proteins encoded by 17 SBP-box genes (Supplementary Table 1) were found. Two genes named *SPL13A* and *SPL13B* in TAIR, which were identified from the genome locus At5g50570 and At5g50670 respectively, shared the same nucleotide and deduced amino acid sequences. Because these two proteins were always related closely with each other, here we just selected *SPL13A* (*AtSPL13* in Supplementary Table 1) for further analysis. Nine alternative splicing genes were found for this family in *Arabidopsis*, the corresponding genome loci were At5g43270, At1g53160, At1g69170, At5g18830, At1g02065, At2g42200, At1g27370, At1g27360 and At5g50570. The gene in At5g43270 (*AtSPL2* in Supplementary Table 1) encoded 3 different proteins as the results of alternative splicing, while others all encoded 2 different proteins. Because the proteins encoded by one alternative splicing gene shared the same sequences, they were constantly related closely with each other in multiple sequences alignment and phylogenetic analysis. Therefore, we only selected one form of proteins encoded by an alternative splicing gene. The gene *AtSPL16* in genome locus At1g76580 may be an alternative splicing gene, too. Its deduced protein contained 489 amino acid residues in TIGR, 810 amino acid residues in TAIR, and 988 amino acid residues in NCBI, respectively. Because only the latter contained the SBP domain, we just selected the protein from NCBI database for further analysis and deduced its CDS sequences based on the genome and protein sequences. The nucleotides, CDS and protein sequences of other SBP-box genes were downloaded from the TIGR database.

According to the annotation information, the SBP-box genes were dispersed on all the *Arabidopsis* chromosomes except for chromosome 4. Seven SBP-box genes were present on chromosome 1, while three SBP-box genes on other chromosomes. According to their position on chromosomes, the genes *AtSPL13A*/*AtSPL13B* and *AtSPL10* /*AtSPL11* were located in two tandem duplications respectively.

3.1.2. SBP-box genes in rice

In 2002, the genomes for *Oryza sativa* ssp. *indica* and *Oryza sativa* ssp. *japonica* were sequenced using shotgun method by Beijing Genomics Institute and Syngenta Company respectively. The International Rice Genome Sequencing Project (IRGSP) published their results of high quality sequences of *Oryza sativa* ssp. *japonica* in 2005. Because the latter one is more accurate and more complete, we only used the *Oryza sativa* ssp. *japonica* sequences from IRGSP for further analysis in this article.

The *OsSPLs* were identified in two steps from the genome of *Oryza sativa* ssp. *japonica* cv *Nipponbare*. The first step involved a blastp search of putative proteins at TIGR database with the SBP domain sequences inferred from *Arabidopsis* and other plants SBP-box proteins as queries. In the second step, each predicted OsSPL protein sequence was confirmed by a Pfam search for the presence of conserved SBP domain. We got 28 proteins which were encoded by 19 genes (Supplementary Table 2) in rice. Five genes were found to be alternative splicing genes. The genome loci were Os01g18850, Os02g04680, Os02g07780, Os06g49010 and Os07g32170. The gene *OsSPL1* in locus Os01g18850 was deduced to encode 3 different proteins, whereas the gene *OsSPL3* (in locus Os02g04680), *OsSPL4* (Os02g07780) and *OsSPL13* (Os07g32170) were deduced to encode 2 proteins respectively, and *OsSPL12* (Os06g49010) to 5 proteins. Because the proteins encoded by one gene shared the same sequences, they were always related closely with each other in further analysis. So we only selected one protein for the alternative splicing genes in multiple sequences alignment and phylogenetic analysis. We renamed all the SBP-box genes in rice based on their order on chromosomes. They were named *OsSPL1*–*OsSPL19* hereafter. After getting the basic information of SBP-box genes in rice, we downloaded the nucleotide sequences in rice genome, CDS and deduced amino acid sequences in TIGR database.

The 19 rice SBP-box genes be distributed on 10 of the 12 rice chromosomes. No SBP-box gene was detected on chromosomes 10 and 12. One SBP-box gene was present on chromosomes 3, 5, 7 and 11, two genes on chromosomes 1, 4 and 9, three genes on chromosomes 2, 6 and 8, respectively. There was no evidence that the SBP-box genes were located in tandem duplications. There was at least one intron in each gene except for the shortest gene *OsSPL19*.

3.2. Phylogenetic analysis of SBP-box family in Arabidopsis and rice

In order to evaluate the evolutionary relationship between *Arabidopsis* and rice SBP-box proteins, and to predict the paralogous and othologous relations among SBP-box proteins in two model plants, the deduced amino acid sequences of the SBP-box genes identified in *Arabidopsis* and *Oryza sativa* ssp. *japonica* were completely aligned. Then a combined phylogenetic tree (Fig. 1) was constructed using software Clustal X 1.83 with neighbor-joining method and bootstrap analysis (1000 reiterations). We divided the SBP-box proteins into 3 major groups based on the phylogenetic tree, where each group contained at least one *Arabidopsis* and rice SBP-box protein. This result indicated that the main characteristics of SBP-box gene family were established before the split of *Arabidopsis* and rice. In order to describe the paralogous and orthologous relations among this family, we divided group A into 6 subgroups and group C into 2 subgroups. As a result, the SBP-box family was divided into 9 subgroups. The bootstrap values for all the subgroups were very high, suggesting that the genes in each subgroup might share a similar origin. Only one pair of orthologous proteins for *Arabidopsis* and rice was identified, i.e., AtSPL7 and OsSPL9 in subgroup A1.4. Six pairs of paralogous proteins in *Arabidopsis* were identified for SBP-box

Please cite this article as: Yang, Z. et al. Comparative study of SBP-box gene family in *Arabidopsis* and rice. Gene (2007), doi:10.1016/j.gene.2007.02.034

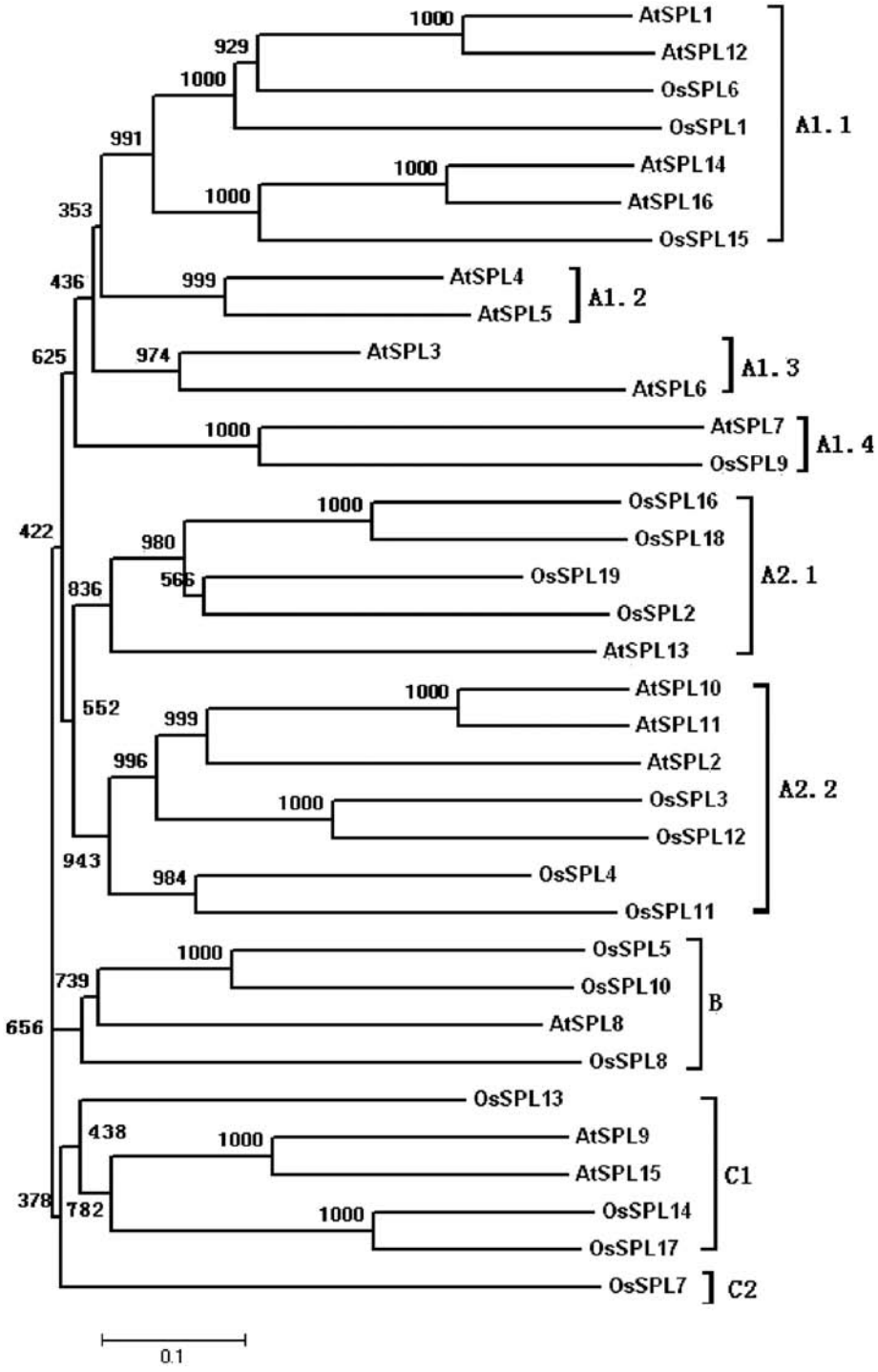

Fig. 1. The phylogenetic tree for the *Arabidopsis* and rice SBP-box proteins. The tree was constructed from a complete alignment of 16 *Arabidopsis* and 19 rice SBP-box proteins by neighbor-joining methods with bootstrapping analysis (1000 reiterations). The groups and subgroups of homologous genes identified are indicated.

family, i.e., AtSPL1 and AtSPL12, AtSPL14 and AtSPL16 in subgroup A1.1; AtSPL4 and AtSPL5 in subgroup A1.2; AtSPL3 and AtSP6 in subgroup A1.3; AtSPL10 and AtSPL11 in subgroup A2.2; AtSPL9 and AtSPL15 in subgroup C1. We also identified 6 pairs of paralogous proteins in rice, which are OsSPL16 and OsSPL18, OsSPL2 and OsSPL19 in subgroup A2.1; OsSPL3 and OsSPL12, OsSPL4 and OsSPL11 in subgroup A2.2; OsSPL5 and OsSPL10 in group B; OsSPL14 and OsSPL17 in subgroup C1. Most proteins in SBP-box family were contained in paralogous pairs (75% for *Arabidopsis* and 63% for rice). This result indicated that most SBP-box genes expanded in a species-specific manner, and probably only a few members originated from the common ancestral genes that existed before the split of *Arabidopsis* and rice. This might be the results of duplications for some genes after the split of two model plants.

Please cite this article as: Yang, Z. et al. Comparative study of SBP-box gene family in *Arabidopsis* and rice. Gene (2007), doi:10.1016/j.gene.2007.02.034

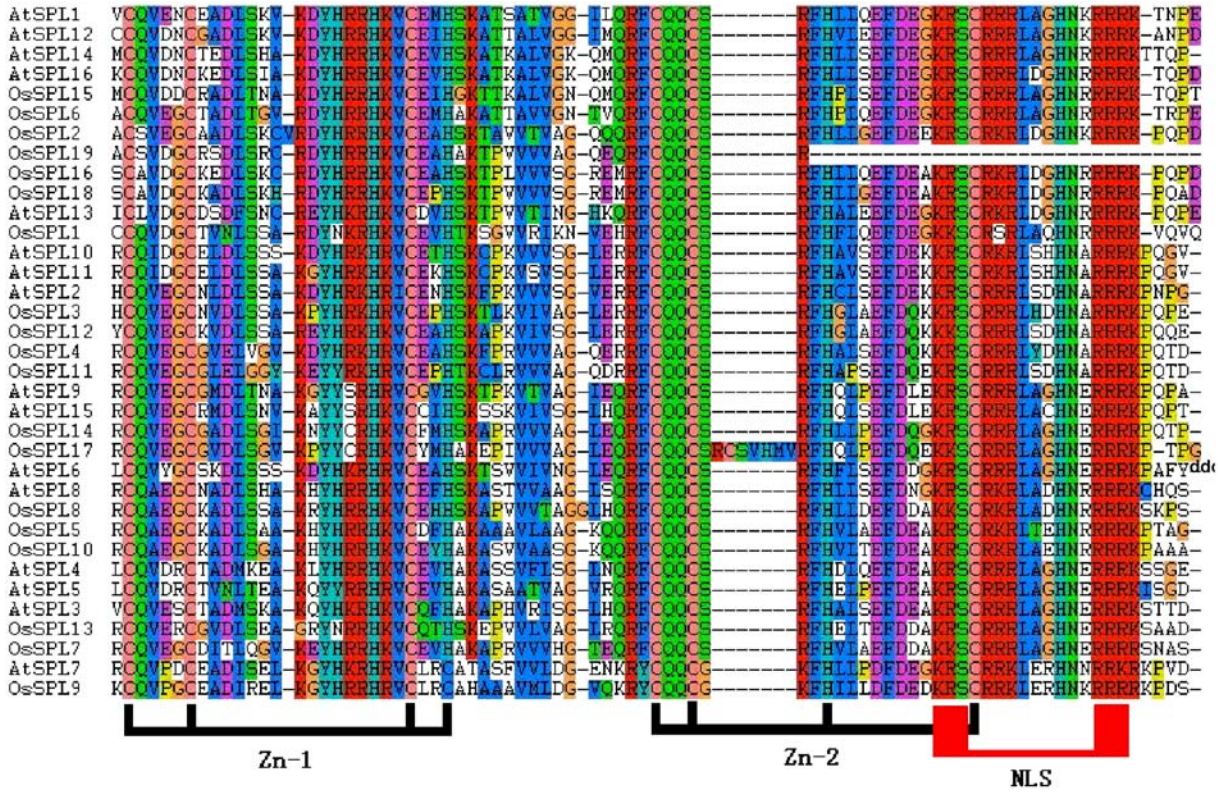

Fig. 2. The SBP domain sequence alignment of the SBP-box proteins in *Arabidopsis* and rice. Multiple sequences alignment was performed using software Clustal X 1.83. The conserved 2 Zn-finger structures and NLS were indicated.

3.3. Conserved sequences in SBP-box proteins

All the members of the SBP-box family contain a conserved SBP domain. We identified the SBP domains contained in all the deduced SBP-box proteins in *Arabidopsis* and rice using Pfam program, then we aligned the domain sequences using software Clustal X 1.83 and found that each SBP domain sequence contained approximately 79 amino acid residues except for OsSPL17 and OsSPL19 in rice (Fig. 2). The structural basis for this sequence-specific binding of DNA are two Zn-finger like structures formed by the coordination of two zinc ions by conserved cysteine and histidine residues (Birkenbihl et al., 2005). The first Zn-finger like structure (Zn-1 in Fig. 2) is Cys_3His-type. The second (Zn-2 in Fig. 2) is $Cys_2HisCys$-type. In addition to their sequence-specific DBD, SBP domain proteins shared a highly conserved bipartite nuclear localization signal (NLS) (Birkenbihl et al., 2005), another feature characteristic for transcription factors, which was located in C-terminus of the SBP domain. It was interesting to find that the His in the first Zn-finger structure was replaced by a Cys residue in AtSPL17 and OsSPL9. The protein OsSPL19 only had the former part of SBP domain sequence, and completely contained only the first Zn-finger structure, because approximately 30 amino acid residues were absent in its C-terminus. There were 7 redundant amino acid residues in the middle part of the SBP domain for the rice protein OsSPL17. In addition to the conserved two Zn-finger structures and NLS segments, we also found that there were only several amino acid residues being replaced for the sequence between the second Zn-finger structure and NLS section, i.e., the sequences between the second Zn-finger structure and NLS section were also conserved.

The identification of motifs for all the *Arabidopsis* and rice SBP-box proteins were performed by the software MEME (Bailey and Elkan, 1994) with default setting except that the maximum number of motifs to find was defined as 40 and the maximum width was setting at 200. We got the sequences of 40 motifs and the distribution of these motifs in SBP-box proteins (Supplementary Fig. 1). Some motifs were found distributed diffusely among SBP-box proteins (e.g. motifs 5, 2, 7, 1 and 6). And all SBP-box proteins contained the motifs 5, 2 and 7. The

Table 1
The *Ka*/*Ks* ratios for all paralogous SBP-box proteins

Paralog pairs	SBP domain	Outside SBP domain
AtSPL1–AtSPL12	0.1367	0.3824
AtSPL14–AtSPL16	0.1674	0.5220
AtSPL4–AtSPL5	0.2041	0.4045
AtSPL3–AtSPL6	0.1873	1.9490
AtSPL10–AtSPL11	0.1815	0.7660
AtSPL9–AtSPL15	0.1966	0.5097
OsSPL16–OsSPL18	0.1345	0.5399
OsSPL2–OsSPL19	0.0998	0.2385
OsSPL3–OsSPL12	0.0860	0.5371
OsSPL5–OsSPL10	0.1580	0.8979
OsSPL4–OsSPL11	0.1191	0.5502
OsSPL14–OsSPL17	0.3055	0.6643

Please cite this article as: Yang, Z. et al. Comparative study of SBP-box gene family in *Arabidopsis* and rice. Gene (2007), doi:10.1016/j.gene.2007.02.034

motifs 1 and 6 were not contained in merely OsSPL19 because a section of sequence was absent in its SBP domain C-terminus. Although many motifs were shared by *Arabidopsis* and rice, there were still species-specific motifs (motifs 20, 23, 29, 31, 34, 35 and 40 in *Arabidopsis* and motifs 25, 27, 30 and 32 in rice). According to Fig. 2, the motifs 5, 2, 7, 1, and 6 were located in the SBP domain.

Based on the subgroups indicated from the phylogenetic tree of SBP-box proteins and motifs through MEME analysis, a schematic distribution of conserved motifs among the defined gene subgroups was constructed (Supplementary Fig. 1). We found: (1) there were similar motifs in each subgroup; (2) the order of motifs in proteins were very similar in each subgroup; (3) Some specific motifs were located in the proteins of specific subgroups, for instance motifs 4, 12, 17 and 22 were specific for subgroup A1.1; (4) Some motifs were only found in specific paralogs (or orthologs), in which motifs 20, 23, 35 were specific for AtSPL10 and AtSPL11, motif 25 for OsSPL17 and OsSPL14, motifs 26, 29 for AtSPL14 and AtSPL16, motifs 27 and 30 for OsSPL16 and OsSPL18, motif 31 for AtSPL9 and AtSPL15, motif 32 for OsSPL5 and OsSPL10, motifs 33 and 38 for AtSPL7 and OsSPL9, motif 34 for AtSPL1 and AtSPL12, motif 40 for AtSPL4 and AtSPL5. These might be related to the function of each subgroup.

The multilevel consensus sequences of motifs identified by MEME (Supplementary Table 3) were used as queries to search for the occurrence of patterns and profiles stored in the PROSITE database. Because motifs 5, 2, 7, 1 and 6 were always found together, their sequences were combined as one for scanning. Two profiles were matched for the joint motifs. One

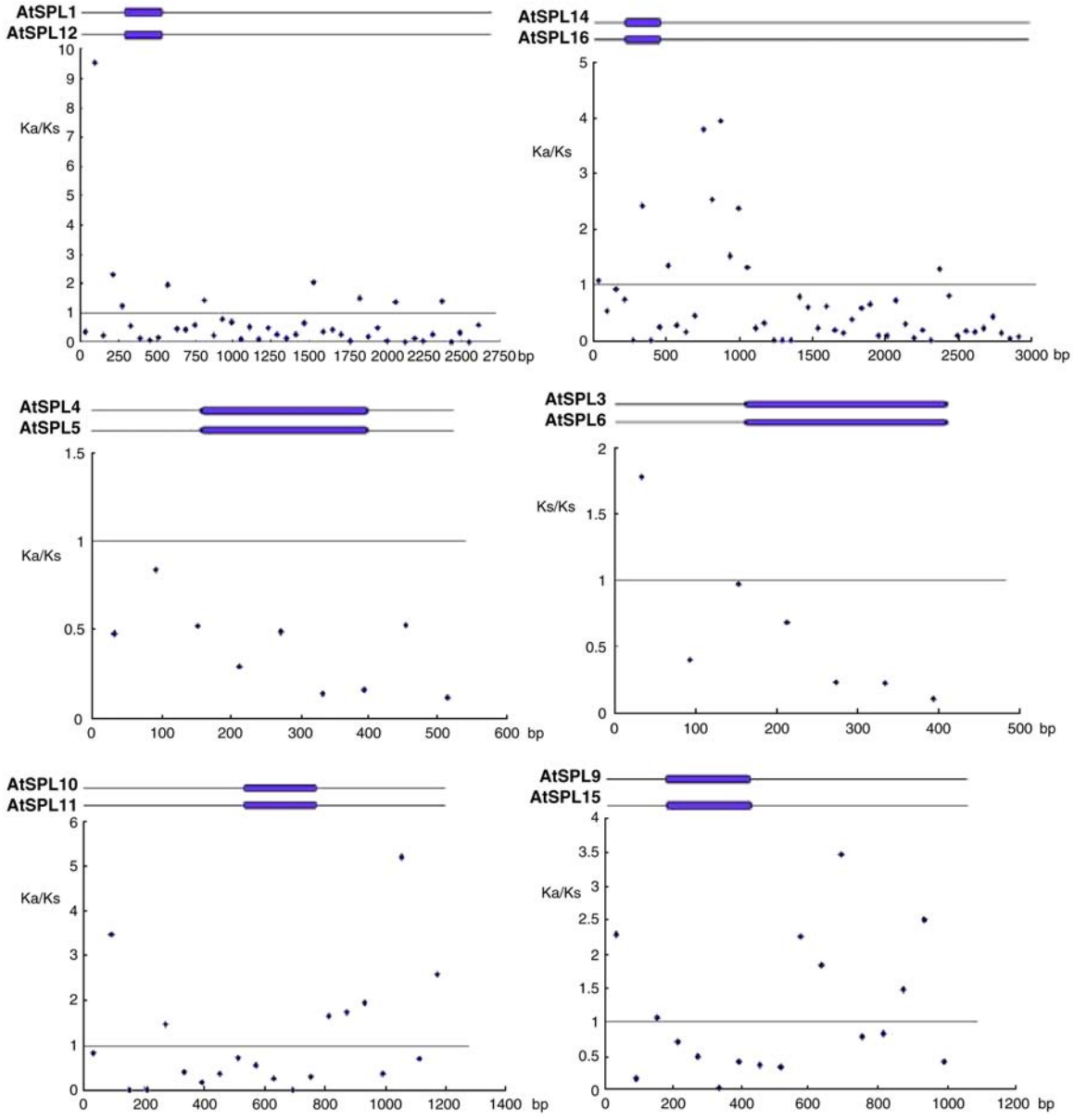

Fig. 3. The *Ka*/*Ks* ratios for *Arabidopsis* SBP-box paralogous proteins with a sliding window of 20 amino acids. The plot shows the *Ka*/*Ks* ratios at various positions for the coding region of *Arabidopsis* SBP-box genes. The thick box represents the SBP domain.

Please cite this article as: Yang, Z. et al. Comparative study of SBP-box gene family in *Arabidopsis* and rice. Gene (2007), doi:10.1016/j.gene.2007.02.034

was the zinc finger SBP-type profile, a SBP domain was identified by Pfam. The other was the bipartite nuclear localization signal profile that located at the C-terminal region of the SBP domain. Ankyrin repeat region circular profile, a protein interaction motif, matched motif 4. This motif was only found in subgroup A1.1. A serine-rich region profile was hit by motif 29, which was found only in three proteins (AtSPL14, AtSPL16 and OsSPL15) of subgroup A1.1. A total of ten PROSITE patterns were identified for all the 40 motifs (a detailed list of them is provided in Supplementary Table 4), while eight of them, i.e., motifs 8, 9, 13, 15, 22, 28, 32, 34 and 39, yielded no hits.

3.4. Driving forces for genetic divergence

To explore whether Darwinian positive selection was involved in driving gene divergence after duplication, the coding regions of six paralogs in *Arabidopsis* and six ones in rice were used to calculate the nonsynonymous/synonynmous substitution ratio (*Ka*/*Ks*). Generally, *Ka*/*Ks* ratio >1 indicates positive selection, a ratio <1 indicates negative or purifying selection and a ratio=1 indicates neutral evolution (Wang et al., 2005b). *Ka*/*Ks* ratios were always less than 1 for SBP domain, suggesting purifying selection on this domain. The *Ka*/*Ks* ratios outside SBP domain were found constantly much higher than the ratios inside SBP domain (Table 1). These indicated that the regions outside SBP domain evolved faster than the SBP domains. These results might be caused by relaxed purifying or positive selection in the regions outside SBP domain. We also calculated the *Ka*/*Ks* ratios for all the paralogs with sliding window of 20 aa (Fig. 3 for *Arabidopsis* and Fig. 4 for rice). The ratios that were higher than 1 were always in the inter-motif regions suggesting these regions had gone through positive selections.

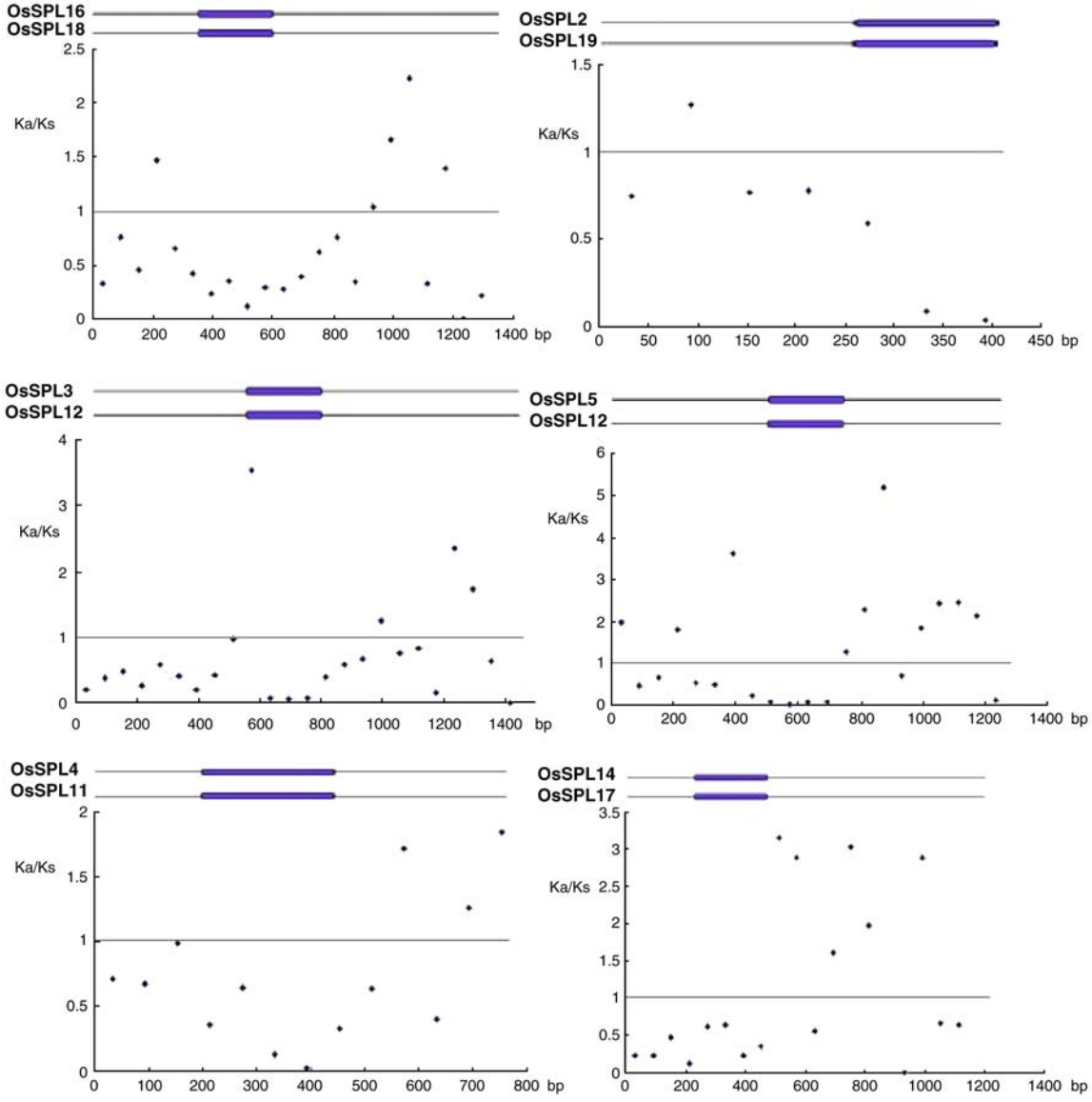

Fig. 4. The *Ka*/*Ks* ratios for rice SBP-box paralogous proteins with a sliding window of 20 amino acids. The plot shows the *Ka*/*Ks* ratios at various positions for the coding region of rice SBP-box genes. The thick box represents the SBP domain.

Please cite this article as: Yang, Z. et al. Comparative study of SBP-box gene family in *Arabidopsis* and rice. Gene (2007), doi:10.1016/j.gene.2007.02.034

Table 2
The distribution for the ESTs of SBP-box genes in *Arabidopsis* and rice

Genes	Number of ESTs	Tissue					
		Callus	Cell Suspension	Seedling	Inflorescence	Rosette	Roots
AtSPL1	30			+	+	+	+
AtSPL2	0						
AtSPL3	16				+	+	+
AtSPL4	10				+		+
AtSPL5	19				+		+
AtSPL6	8				+		+
AtSPL7	17				+	+	+
AtSPL8	5				+		+
AtSPL9	18				+	+	+
AtSPL10	14				+	+	+
AtSPL11	41	+	+		+		+
AtSPL12	17				+	+	
AtSPL13	23				+	+	+
AtSPL14	13			+	+	+	+
AtSPL15	12				+	+	+
AtSPL16	1		+		+		+
OsSPL1	50	+			+	+	+
OsSPL2	2				+		
OsSPL3	28	+			+	+	+
OsSPL4	1				+		
OsSPL5	1				+		
OsSPL6	50	+			+	+	+
OsSPL7	3				+		
OsSPL8	4				+	+	
OsSPL9	45	+			+	+	+
OsSPL10	1				+		
OsSPL11	3				+		
OsSPL12	49	+			+		
OsSPL13	8				+		
OsSPL14	9	+			+		
OsSPL15	50	+			+	+	
OsSPL16	7				+		
OsSPL17	2	+			+		
OsSPL18	0						
OsSPL19	0						

3.5. The expression analysis of SBP-box genes in Arabidopsis and rice

Expressed sequence tags (ESTs) provide a useful means of studying mRNA expression profiles (digital northerns) (Ohlrogge and Benning, 2000). The frequency of ESTs or cDNAs available in different database was considered as a useful tool for preliminary analysis of gene expression (Adams et al., 1995). At the time of analysis, the Genbank EST database contained 622,973 ESTs for *Arabidopsis* and 1,188,565 ESTs for rice. In order to get a better picture of the tissue-specific expression of *Arabidopsis* and rice SBP-box genes, we divided the ESTs into 6 tissue categories according to GENEVESTIGATOR (Zimmermann et al., 2004). The database was searched with the CDS sequences of each SBP-box genes using blastn program, and the resulting accessions were classified according to the source tissue from which the cDNA libraries derived (Table 2).

We could not find evidence of the expression for genes *AtSPL2* in *Arabidopsis* and *OsSPL18*, *OsSPL19* in rice. This might be results of characteristics of specially temporal and spacial expression pattern for genes. The EST database did not contain the EST resource for these genes. On the other hand, these genes might be the pseudogenes, especially for *OsSPL19*, because it contained an incomplete SBP domain, which might be the results of transposon mutagenesis and lost of its functional section. All the other SBP-box genes could be expressed in inflorescence, which were consistent with the function of SBP-box genes for controlling flower development. In *Arabidopsis*, most genes were expressed mainly in the tissue of inflorescence, roots, and rosette. In addition to inflorescence, there were nearly half of the rice SBP-box genes expressing in the tissue of callus, while only *AtSPL11* in *Arabidopsis* was found to express in callus, indicating that the SBP-box genes had gone through partially functional differentiation after the split of *Arabidopsis* and rice.

4. Discussion

Rice and *Arabidopsis*, being widely accepted as the model plants of monocotyledon and dicotyledon respectively, diverged from a common ancestor about 200 million years ago (Wolfe et al., 1989). Fundamental differences between the *Arabidopsis*

Please cite this article as: Yang, Z. et al. Comparative study of SBP-box gene family in *Arabidopsis* and rice. Gene (2007), doi:10.1016/j.gene.2007.02.034

and rice genomes include their size and gene content, the rice genome being larger and containing more genes than that of *Arabidopsis* (*Arabidopsis*: about 146 Mbp and 26207 genes (Haas et al., 2005), rice: about 389 Mbp and 31544 genes (IRGSP, 2005)). The genomes of two plants shared very limited synteny: about 71% of the deduced genes in rice had a putative homologue in *Arabidopsis*. In a reciprocal analysis, 90% of the *Arabidopsis* proteins had a putative homologue in the predicted rice proteome (IRGSP, 2005). In addition to the complete genomic sequences of *Arabidopsis thaliana* and rice, some institutes had done a lot to annotate the two genomes (Wortman et al., 2003; Haas et al., 2005; Yuan et al., 2005; Ohyanagi et al., 2006), which laid a foundation for identifying all the genes in *Arabidopsis* and rice. To compare the members in a gene family from different species based on the information of annotation of *Arabidopsis* and rice genomes had become an important method to illuminate the characteristics of conservation, divergence and evolution for monocotyledonous and dicotyledonous plants (Lijavetzky et al., 2003; Shiu et al., 2004; Martinez et al., 2005; Feng et al., 2006).

SBP-box gene family belongs to the type of plant-specific zinc finger protein genes, which encode plant-specific transcription factor, and the proteins encoded by SBP-box genes could specifically bind to promoters of the floral meristem identity gene *SQUAMOSA* and its orthologous genes. Some researchers had provided the evidence that SBP-box genes were present in *Arabidopsis* (Cardon et al., 1997; Unte et al., 2003; Birkenbihl et al., 2005; Stone et al., 2005), maize (Becraft et al., 1990; Wang et al., 2005a), *Antirrhinum majus* (Klein et al., 1996), rice (Shao et al., 1999) and silver birch (Lannenpaa et al., 2004). In this article, we presented the comparative phylogenetic analysis of SBP-box gene family from *Arabidopsis* and rice.

There were 17 SBP-box genes in *Arabidopsis*. Among them, *AtSPL13A* and *AtSPL13B*, *AtSPL10* and *AtSPL11* were located in 2 tandem repeats, where the former pair of genes shared the same nucleotide and amino acid sequences. There were 19 SBP-box genes in rice. The protein products of all these genes contained a highly conserved SBP domain. The plant-specific SBP domains contained about 79 amino acid residues except for OsSPL19 and OsSPL17, and they were highly conserved in all SBP-box proteins from *Arabidopsis* and rice. The structural basis for this sequence-specific binding of DNA are two Zn-finger like structures formed by the coordination of two zinc ions and a bipartite nuclear localization signal section.

A phylogenetic tree was constructed using the amino acid sequences encoded by 35 SBP-box genes from *Arabidopsis* and rice. According to the phylogenetic tree we divided the SBP-box proteins into 3 major groups, and each group contained at least one *Arabidopsis* and rice SBP-box proteins, which indicated that the main characteristics of this family were formed before the split of monocotyledonous and dicotyledonous plants. The main objective of this phylogenetic study was to identify putative orthologous and paralogous SBP-box genes. Paralogs usually display different functions, while orthologs may retain the same function. In 35 SBP-box proteins, we found only one pair of orthologous proteins, but 6 pairs of paralogous proteins in both *Arabidopsis* and rice respectively. The fact that the genes in paralogs accounted for the most of the family indicated that the two model plants might have gone through several duplication events after split and most SBP-box genes in *Arabidopsis* and rice expanded in a species-specific manner. This type of divergence between a monocot and a dicot species had been observed for other gene families as well (Bai et al., 2002; Zhang et al., 2005; Jain et al., 2006).

We divided the members of this family into 9 subgroups, and identified the motifs for the proteins using program MEME. The motif structures of SBP-box proteins were mostly conserved between *Arabidopsis* and rice, indicating that motif organizations were mostly established before their divergence. There were similar motifs and the orders of motifs in the proteins for each subgroup were almost identical. These indicated that the genes in same subgroup might have similar functions in plant development. The *Arabidopsis* proteins in subgroups A1.1 were the largest proteins. Based on the reports of Cardon and Schmid (Cardon et al., 1999; Schmid et al., 2003), the *Arabidopsis* genes encoding these proteins were expressed constitutively, while the mid-sized and small genes were up-regulated mainly in flower development (Birkenbihl et al., 2005). According to the phylogenetic tree, the genes in rice belonged to the subgroup A1.1 (gene *OsSPL1*, *OsSPL6* and *OsSPL15*) might be expressed constitutively, since these genes were very large and shared similar sequences with *Arabidopsis* genes in A1.1. Other genes in rice were mid or small size. The functions of these genes might be focused on up-regulation in flower development. The *Arabidopsis* proteins AtSPL14, AtSPL16 and the rice protein OsSPL15 were located in the same clade in the phylogenetic tree, and they shared the ankyrin repeat region circular profile and serine-rich region profile. The motif structures identified by MEME among them were quite similar. *AtSPL14* had recently been characterized as conferring resistance to the programmed cell death (PCD)-inducing fungal toxin fumonisin B1 (FB1) (Stone et al., 2005). The *Arabidopsis* gene *AtSPL16* and the rice gene *OsSPL15* may have a same function with *AtSPL14*. We also found that species-specific and paralogs-specific motifs. These results suggest that motif acquisition/divergence had continued to occur in *Arabidopsis* and rice after split.

In *Arabidopsis* and rice, the SBP domain had gone through purifying selection, while some regions outside SBP domain had gone through positive selection or relaxed purifying selection. Positive selection is one of the major driving forces for the emergence of new functions in proteins after gene duplication. In our analysis, the *Ka*/*Ks* ratios that were higher than 1 were always in the inter-motif regions. These may be related to the emergence of new functions for SBP-box genes. Such positive selection was also found in RLK (Shiu et al., 2004) and Dof (Yang et al., 2006) gene families in *Arabidopsis* and rice.

We did not find the evidences of expression for genes *AtSPL2*, *OsSPL18* and *OsSPL19*. This might be caused by specially temporal and spacial expression patterns. Or, possibly, these genes are pseudogenes. For instance, we did not find the expressional EST for *OsSPL19*, meanwhile the SBP domain in the protein encoded by this gene was not integrated. The EST

Please cite this article as: Yang, Z. et al. Comparative study of SBP-box gene family in *Arabidopsis* and rice. Gene (2007), doi:10.1016/j.gene.2007.02.034

sequences were identified for most genes in this research and most of them were found mainly expressed in inflorescence. This is consistent with the function of SBP-box genes for controlling flower development. In addition to flowers, there were nearly half of the rice SBP-box genes expressing in the tissue of callus, indicating that these genes were related to the process of dedifferentiation and plant architecture. But only *AtSPL11* in *Arabidopsis* was found to express in callus. These suggested that the SBP-box genes had gone through partially functional differentiation after the split of *Arabidopsis* and rice.

In this study, we carried out a comprehensive comparative study about the SBP-box gene family in *Arabidopsis* and rice. The results deserve further experimental verification and comparison. Currently, there are around 41 plant genomic sequencing projects in progress, but only *Arabidopsis* and rice provide complete and well-annotated sequence. We additionally searched two uncompleted genomes, i.e., *Populus trichocarpa* and *Chlamydomonas reinhardtii*. Their draft sequences and partial annotations are accessible at the DOE JGI website (http://www.jgi.doe.gov/). A cluster of SBP-box genes were detected in both genomes (Supplementary Tables 5 and 6). We also detected a total of 32 pieces of protein in 12 species got by searching NCBI database (Supplementary Table 7). But up to now, it was hard to decide which sequence should be included to the phylogeny or motif analysis as we found most matches were not in full-length. Beside, such search was not exhaustive as the sequences were incomplete. It was important, however, to incorporate more fully sequenced genomes (possibly *Populus* since it had been under assembly) in the future study. Large data set allow studying the evolution of SBP-box genes in greater detail. It may help infer the ancestral groups of the genes and hopefully ancestor–descendant relationships could be better illustrated.

Acknowledgements

The authors are grateful to the editor and two anonymous reviewers for their helpful comments and criticisms. The authors also thank Dr. Emmanuel Paradis for a thorough reading of the original manuscript. This work was supported by the National Basic Research Program of China (grant no. 2006CB101700), the National High-tech R&D Program (grant no. 2006AA10Z165), the National Natural Science Foundation of China (grant no. 30370758) and Program for New Century Excellent Talents in University.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.gene.2007.02.034.

References

Adams, M.D., et al., 1995. Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. Nature 377, 3–174.

Bai, J., et al., 2002. Diversity in nucleotide binding site-leucine-rich repeat genes in cereals. Genome Res. 12, 1871–1884.

Bailey, T.L., Elkan, C., 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology. AAAI Press, pp. 28–36.

Becraft, P.W., Bongard-Pierce, D.K., Sylvester, A.W., Poethig, R.S., Freeling, M., 1990. The liguleless-1 gene acts tissue specifically in maize leaf development. Dev. Biol. 141, 220–232.

Birkenbihl, R.P., Jach, G., Saedler, H., Huijser, P., 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.H., Hohmann, S., Nettesheim, K., Saedler, H., Huijser, P., 1997. Functional analysis of the *Arabidopsis thaliana* SBP-box gene SPL3, a novel gene involved in the floral transition. Plant J. 12, 367–377.

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

Comeron, J.M., 1999. K-estimator, calculation of the number of nucleotide substitutions per site and the confidence intervals. Bioinformatics 15, 763–764.

Feng, Y., Liu, Q., Xue, Q., 2006. Comparative study of rice and *Arabidopsis* actin-depolymerizing factors gene families. J. Plant Physiol. 163, 69–79.

Goff, S.A., et al., 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). Science 296, 92–100.

Haas, B.J., et al., 2005. Complete reannotation of the *Arabidopsis* genome, methods, tools, protocols and the final release. BMC Biol. 3, 7.

IRGSP, 2005. The map-based sequence of the rice genome. Nature 436, 793–800.

Jain, M., Tyagi, A.K., Khurana, J.P., 2006. Genome-wide analysis, evolutionary expansion, and expression of early auxin-responsive SAUR gene family in rice (*Oryza sativa*). Genomics 88, 360–371.

Klein, J., Saedler, H., Huijser, P., 1996. A new family of DNA binding proteins includes putative transcriptional regulators of the *Antirrhinum majus* floral meristem identity gene *SQUAMOSA*. Mol. Gen. Genet. 250, 7–16.

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

Lijavetzky, D., Carbonero, P., Vicente-Carbajosa, J., 2003. Genome-wide comparative phylogenetic analysis of the rice and *Arabidopsis* Dof gene families. BMC Evol. Biol. 3, 17.

Martinez, M., Abraham, Z., Carbonero, P., Diaz, I., 2005. Comparative phylogenetic analysis of cystatin gene families from *arabidopsis*, rice and barley. Mol. Genet. Genomics 273, 423–432.

Ohlrogge, J., Benning, C., 2000. Unraveling plant metabolism by EST analysis. Curr. Opin. Plant Biol. 3, 224–228.

Ohyanagi, H., et al., 2006. The Rice Annotation Project Database (RAP-DB), hub for *Oryza sativa* ssp. *japonica* genome information. Nucleic Acids Res. 34, 741–744.

Riechmann, J.L., et al., 2000. *Arabidopsis* transcription factors, genome-wide comparative analysis among eukaryotes. Science 290, 2105–2110.

Saitou, N., Nei, M., 1987. The neighbor-joining method, a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4, 406–425.

Schmid, M., et al., 2003. Dissection of floral induction pathways using global expression analysis. Development 130, 6001–6012.

Shao, C.X., Takeda, Y., Hatano, S., Matsuoka, M., Hirano, H.-Y., 1999. Rice genes encoding the SBP domain protein, which is a new type of transcription factor controlling plant development. Rice Genet. Newsl. 16, 114.

Shiu, S.H., Karlowski, W.M., Pan, R., Tzeng, Y.H., Mayer, K.F., Li, W.H., 2004. Comparative analysis of the receptor-like kinase family in *Arabidopsis* and rice. Plant Cell 16, 1220–1234.

Sonnhammer, E.L., Eddy, S.R., Durbin, R., 1997. Pfam, a comprehensive database of protein domain families based on seed alignments. Proteins 28, 405–420.

Stone, J.M., Liang, X., Nekl, E.R., Stiers, J.J., 2005. *Arabidopsis* AtSPL14, a plant-specific SBP-domain transcription factor, participates in plant development and sensitivity to fumonisin B1. Plant J. 41, 744–754.

Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F., Higgins, D.G., 1997. The CLUSTAL_X windows interface, flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. 25, 4876–4882.

Unte, U.S., et al., 2003. SPL8, an SBP-box gene that affects pollen sac development in *Arabidopsis*. Plant Cell 15, 1009–1019.

Wang, H., et al., 2005a. The origin of the naked grains of maize. Nature 436, 714–719.

Wang, W., et al., 2005b. Origin and evolution of new exons in rodents. Genome Res. 15, 1258–1264.

Wolfe, K.H., Gouy, M., Yang, Y.W., Sharp, P.M., Li, W.H., 1989. Date of the monocot-dicot divergence estimated from chloroplast DNA sequence data. Proc. Natl. Acad. Sci. U. S. A. 86, 6201–6205.

Wortman, J.R., et al., 2003. Annotation of the *Arabidopsis* genome. Plant Physiol. 132, 461–468.

Xiong, Y., Liu, T., Tian, C., Sun, S., Li, J., Chen, M., 2005. Transcription factors in rice, a genome-wide comparative analysis between monocots and eudicots. Plant Mol. Biol. 59, 191–203.

Yamasaki, K., et al., 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, X., Tuskan, G.A., Cheng, M.Z., 2006. Divergence of the Dof gene families in poplar, *Arabidopsis* and rice suggests multiple modes of gene evolution after duplication. Plant Physiol. 142, 820–830.

Yu, J., et al., 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). Science 296, 79–92.

Yuan, Q., et al., 2005. The institute for genomic research Osa1 rice genome annotation database. Plant Physiol. 138, 18–26.

Zhai, Y., Tchieu, J., Saier Jr., M.H., 2002. A web-based Tree View (TV) program for the visualization of phylogenetic trees. J. Mol. Microbiol. Biotechnol. 4, 69–70.

Zhang, S., et al., 2005. Evolutionary expansion, gene structure, and expression of the rice wall-associated kinase gene family. Plant Physiol. 139, 1107–1124.

Zimmermann, P., Hirsch-Hoffmann, M., Hennig, L., Gruissem, W., 2004. GENEVESTIGATOR. *Arabidopsis* microarray database and analysis toolbox. Plant Physiol. 136, 2621–2632.

Please cite this article as: Yang, Z. et al. Comparative study of SBP-box gene family in *Arabidopsis* and rice. Gene (2007), doi:10.1016/j.gene.2007.02.034