Mol Gen Genomics (2006) 276:478–494
DOI 10.1007/s00438-006-0159-z

ORIGINAL PAPER

Putative fasciclin-like arabinogalactan-proteins (FLA) in wheat (*Triticum aestivum*) and rice (*Oryza sativa*): identification and bioinformatic analyses

Ahmed Faik · Jaouad Abouzouhair · Fathey Sarhan

Received: 9 June 2006 / Accepted: 7 August 2006 / Published online: 31 August 2006

Abstract Putative plant adhesion molecules include arabinogalactan-proteins having fasciclin-like domains. In animal, fasciclin proteins participate in cell adhesion and communication. However, the molecular basis of interactions in plants is still unknown and none of these domains have been characterized in cereals. This work reports the characterization of 34 wheat (*Triticum aestivum*) and 24 rice (*Oryza sativa*) Fasciclin-Like Arabinogalactan-proteins (FLAs). Bioinformatics analyses show that cereal FLAs share structural characteristics with known Arabidopsis FLAs including arabinogalactan-protein and fasciclin conserved domains. At least 70% of the wheat and rice FLAs are predicted to be glycosylphosphatidylinositol-anchored to the plasma membranes. Expression analyses determined from the relative abundance of ESTs in the publicly available wheat EST databases and from RNA gel blots indicate that most of these genes are weakly expressed and found mainly in seeds and roots. Furthermore, most wheat genes were down regulated by abiotic stresses except for *TaFLA9* and *12* where cold treatment induces their expression in roots. Plant fasciclin-like domains were predicted to have 3-D homology with FAS1 domain of the fasciclin I insect neural cell adhesion molecule with an estimated precision above 70%. The structural analysis shows that negatively charged amino acids are concentrated along the β1-α3-α4-β2 edges, while the positively charged amino acids are concentrated on the back side of the folds. This highly charged surface distribution could provide a way of mediating protein–protein interactions via electrostatic forces similar to many other adhesion molecules. The identification of wheat FLAs will facilitate studying their function in plant growth and development and their role in stress response.

Keywords Arabinogalactan-proteins · Fasciclin-like domain · Glycosylphosphatidylinositol anchor · Bioinformatics · Cereals · Wheat · Rice

Communicated by R. Hagemann.

Nucleotide sequence data reported are available in the DBJ/EMBL/GenBank databases under the accession numbers: *TaFLA1*, DQ872374; *TaFLA2*, DQ872375; *TaFLA3*, DQ872376; *TaFLA4*, DQ872377; *TaFLA5*, DQ872378; *TaFLA6*, DQ872379; *TaFLA7*, DQ872380; *TaFLA8*, DQ872381; *TaFLA9*, DQ872382; *TaFLA10*, DQ872383; *TaFLA11*, DQ872384; *TaFLA12*, DQ872385; *TaFLA13*, DQ872386; *TaFLA14*, DQ872387; *TaFLA15*, DQ872388; *TaFLA16*, DQ872389; *TaFLA17*, DQ872390; *TaFLA18*, DQ872391; *TaFLA19*, DQ872392; *TaFLA20*, DQ872393; *TaFLA21*, DQ872394; *TaFLA22*, DQ872395; *TaFLA23*, DQ872396; *TaFLA24*, DQ872397; *TaFLA25*, DQ872398; *TaFL26*, DQ872399; *TaFLA27*, DQ872400; *TaFLA28*, DQ872401; *TaFLA29*, DQ872402; *TaFLA30*, DQ872403; *TaFLA31*, DQ872404; *TaAGP1*, DQ872405; *TaFLA33*, DQ872406; TaFLA34, DQ872407. If requested the database will withhold release of data until publication.

A. Faik (✉)
Department of Environmental and Plant Biology,
Ohio University, Porter Hall, Athens, OH 45701, USA
e-mail: faik@ohio.edu

J. Abouzouhair · F. Sarhan
Département des sciences biologiques, Université du Québec à Montréal (UQAM), 141 Président-Kennedy, Montreal, QC H2X 3Y5, Canada

Introduction

Plant development and adaptation to continuous environmental changes require that cells undergo profound

structural and metabolic modifications. The plant cell wall (CW) and its interaction with the plasma membrane (PM), is thought to play crucial regulatory roles during these developmental processes. CW–PM interactions are dynamic, and involve various types of proteins such as receptor kinases, arabinogalactan-proteins (AGPs), cellulose synthase complex (rosettes), and wall-associated kinases (WAKs). As part of a study on wheat endosperm development, we sought to characterize putative wheat fasciclin-like arabinogalactan-proteins (TaFLAs) and investigate their possible physiological function in plant development and in response to various stresses. Such putative adhesion molecules are not characterized in cereals. The hydroxyproline (Hyp)-rich glycoproteins (HRGPs) superfamily consists of three main subfamilies, namely arabinogalactan-proteins (AGPs), extensins (EXTs), and proline-rich proteins (PRPs). HRGPs are characterized by a backbone protein rich in Hyp with varying degrees of *O*-glycosylation depending upon the subfamily (up to 99% of M_r) (Nothnagel 1997; Bacic et al. 2000; Showalter 2001). HRGPs share common signatures that consist of blocks of Hyp residues organized in clusters called glycomodules, in which most Hyp residues usually bear either a large arabinogalactan (AG) polysaccharide or small non-branched arabinooligosaccharides. In the case of AG polymers, Hyp residues are organized in noncontiguous clusters (as in AGPs), however, in the case of arabinooligosaccharides, Hyp residues are organized in contiguous clusters as in EXTs and PRPs (Tan et al. 2003). In addition to the AG glycomodules, an AGP molecule may contain other conserved domains such as fasciclin-like domains similar to the ones identified in fruit flies (Elkins et al. 1990), or Lys-rich domains (Gaspar et al. 2004). Although an excellent bioinformatics strategy to identify AGPs has been developed in Arabidopsis genome (Schultz et al. 2002); there is currently an urgent need for an exhaustive bioinformatics analysis and characterization of AGPs in genomes from commercially important crop plants such rice, wheat and maize. Such an analysis may lead to a better understanding of the physiological functions of these proteins. In Arabidopsis, 21 genes encoding FLAs (AtFLAs) have been identified and were classified into four groups (Johnson et al. 2003). The AtFLA genes are expressed in various plant organs, in various tissues such as xylem, stylar tissue, and in suspension cultured cells (Showalter 2001; Johnson et al. 2003; Lafarguette et al. 2004; Dahiya et al. 2006). Structurally, a fasciclin domain, as defined by the consensus "smart00554" in SMART databases (Simple Modular Architecture Research Tool at http://www.smart.embl. de; http://www.smart.ox.ac.uk), is characterized by two highly conserved sequences regions called H1 and H2 (~10 amino acids long each) with common short conserved peptide motifs, namely Val-Phe-Pro-X-X-X-Pro and [Phe/Tyr]-His motifs, where X can be any amino acid (Kawamoto et al. 1998; Johnson et al. 2003). Considering the number of AtFLA proteins and their structural diversity (Lafarguette et al. 2004; Gaspar et al. 2004), it is expected that they may have diverse function during plant development and adaptation. For example, the salt overly sensitive five mutant in Arabidopsis (*sos5*) has a point mutation in the second fasciclin domain of *AtFLA4* gene, resulting in plants with thinner cell walls, abnormal swollen cells at the root tip, and an increase in salt sensitivity (Shi et al. 2003). More recent studies (Lafarguette et al. 2004; Brown et al. 2005; Persson et al. 2005; Dahiya et al. 2006) showed that *AtFLA11* gene and its homolog in *Zinnia elegans* and poplar are implicated in secondary wall formation. It is clear that FLA proteins play crucial roles in plant development, however the mechanisms by which these proteins achieve their functions are not known. To date, the adhesion function (physical interaction) of plant fasciclin-like domains has not been demonstrated as they have been in fruit fly Fas1 and in the immunoglobulin super-family (Wang et al. 1999; Kim et al. 2000, 2002). It has been postulated that the dual presence of fasciclin-like domains associated with glycosylphosphatidylinositol (GPI)-anchor supports the involvement of plant FLA in cell adhesion and signaling. However, we still do not know by which mechanism the interaction may occur. This lack of progress in understanding adhesion mechanism of plant fasciclin-like domains comes mostly from the lack of structural studies of their domains. Recent advances in computer prediction algorithms and the availability of the crystal structure model of FAS1 domain 4 of *Drosophila* fasciclin (Coult et al. 2003), will open a new way to investigate the importance of certain regions of plant fasciclin-like domains in adhesion mechanism. In an effort to understand the function of FLAs in cereal development, we initiated this study that focuses on the identification, cloning, and bioinformatics analyses of putative FLAs in wheat, a commercially important *Triticeae* grain crop. A java script was developed that allowed us to identify wheat ESTs with similarities to Arabidopsis FLA proteins in the public databases and then assemble them into unique contigs, by using bioinformatics and transcriptomic approaches in association with currently available genomic resources. Using this strategy along with PCR-based methodology, we cloned 34 full-length putative *TaFLA* genes. Structural predictions showed that plant fasciclin-like domain has fold homology to FAS1 with an estimated precision of

90%, and the domain consists of a seven-stranded wedge similar to FAS1. This study may establish the foundation for further detailed investigations on the role of each particular TaFLA protein in plant development and environmental adaptation. The cloning of putative *TaFLA* genes in addition to the mining of rice homologues allowed us to perform extensive comparison of these proteins with Arabidopsis proteins. Inventory of wheat *FLA* transcripts in the TIGR EST database indicates that most of *TaFLA* genes are expressed in reproductive organs and roots. In addition, RNA gel blot analysis indicates that at least two of these wheat genes are up regulated by cold treatment.

Materials and methods

Plant material and growth conditions

Winter wheat (*Triticum aestivum* L. cv Norstar, LT_{50} of −19°C) was used in this work. Plants were grown in water-saturated vermiculite for 7 days at 25°C/20°C under a relative humidity of 70% and a 15 h photoperiod. After this period plants were treated with various stresses as described earlier (Danyluk et al. 1998). Control plants were maintained under the same conditions of light and temperature. Cold acclimation was carried out by subjecting germinated seedlings to a temperature of 4°C ± 1°C with a 12 h photoperiod for various periods of time as specified for each experiment. Seedlings were watered with a nutrient solution (20:20:20; N:P:K). Plant tissues were immediately frozen in liquid nitrogen after harvesting and stored at −80°C until use. Heat shock was performed by incubating seedlings at 40°C for 3 h. Salt-stressed plants were obtained by incubating seedlings for 24 h in a nutrient solution containing 500 mM NaCl. Abscisic acid (ABA)–treated plants were obtained by incubating seedlings for 18 h in a nutrient solution containing 10^{-5} M ABA and concomitantly applying a foliar spray containing 10^{-4} M ABA in 0.02% (v/v) Tween 20. Polyethylene glycol-induced osmotic stress was performed by removing seedlings from vermiculite and transferring them to a solution containing 70% (w/v) of polyethylene glycol (average molecular weight of 8,000) for 48 h.

Identification of TaFLA-like AGP ESTs using bioinformatics and transcriptomic approaches

Bioinformatic and transcriptomic approaches were used to screen wheat (*T. aestivum*) EST databases publicly available. An In-house java script was developed to perform several bioinformatics steps yielding a final list of unique wheat sequences with putative fasciclin-like domains. The first step consists of the BLAST search (http://www.ncbi.nlm.nih.gov/Ftp/index.html), to identify and collect hits with similarity ($E < 0.05$) to each of *Arabidopsis thaliana* FLAs (Johnson et al. 2003). Then the script removes the redundancy (identical hits names) from this first crude list of hits before making an assembly of putative wheat contigs using CAP3 program (Huang and Madan 1999). Short sequences (less than 300 pb) with low quality nucleotide sequences (~10–30% NNNs) were also discarded before assembling the contigs. Using clustalW program, alignments of these wheat unique sequences with their closest homologs from rice and Arabidopsis were performed to identify the non-overlapping contigs that might correspond to different regions of the same wheat gene. This last manual step reduced the number of unique genes and generates a final list of putative wheat FLA-like AGP (*TaFLA*) genes. The search in the initial step was first performed using AtFLA nucleic acid gene sequences as query against nucleic acid databases (BlastN), but later we found that TBlastN program (AtFLA proteins vs. dynamically translated EST databases) was more effective in detecting hits. Because of the presence of repetitive motifs in these proteins, a low complexity filter was not used. Multiple sequence alignment of these protein homologs was generated using ClustalW program. Simultaneously, the publicly available clones of the putative *TaFLA* genes found were then fully sequenced at least three times to eliminate errors from sequencing procedures. The comparison of *TaFLA* genes with their closest homologs, allowed us to accurately predict the start (ATG) and stop codons. The remaining putative *TaFLA* genes that did not have clones available (~50% of the total list) were cloned using PCR-based cloning strategy.

PCR-based cloning of wheat *FLA*-like genes

Gene-specific primers were designed for these genes using Primer3 program (http://www.frodo.wi.mit.edu/cgi-bin/primer3/primer3_www.cgi) that were used to PCR amplify cDNA clones from 7 cDNA libraries (in pCMVSPORT6 and are directional) made from various wheat tissues and developmental stages. The reverse and forward primers used with each of the gene-specific primers are 5′-AGATCCCAAGCTAGCAGTTTTCCCAGTCACGA-3′and 5′-GAGCGGATAACAATTTCACACAGGAAACAGCTATGA-3′, respectively. 50 µl PCR reactions contained 0.6 µM

forward gene-specific primer, 0.3 μM of the same reverse primer, 100 ng DNA, 300 μM four dNTPs, 1 U Pfx polymerase as indicated in Platinum Pfx DNA Polymerase kit manual (Invitrogen). PCR conditions were as follows: an initial 2 min step at 94°C to activate Pfx is followed by eight cycles of 20 s denaturing at 94°C, 20 s annealing at 72°C, and 150 s extension at 72°C. The annealing temperature was lowered by 1°C each cycle using down touch function. At the end of these eight cycles the PCR reaction is extended for another 44 cycles of 20 s denaturing step at 94°C, 20 s annealing step at 64°C, and 150 s extension step at 68°C. The final step is conducted at 72°C for 10 min. PCR products were cloned into PCR4 Blunt-TOPO vector using zero blunt TOPO PCR cloning kit (Invitrogen) and then introduced into chemically competent bacteria cells (one shot cells).

Computer-based analysis of protein sequences

Several prediction algorithms are currently available for public use through well-organized websites. These websites are listed as follows: putative fasciclin domains were identified using two websites. The first allows the search the Conserved Domain Database at NCBI (http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi) (Marchler-Bauer and Bryant 2004). This algorithm did not identify the FLA-like domain in six wheat and three rice proteins. Therefore, a motif scan algorithm allowing the search of Prosite and Pfam database of motifs (http://www.myhits.isb-sib.ch/cgi-bin/motif_scan) was used to analyze these proteins. This program is currently the best in predicting motifs that have low homology. For the other proteins, the motifs were extracted manually.

To determine the N-terminal signal sequence for targeting the protein to the secretory pathway, the deduced amino acid sequence of putative wheat and rice FLA proteins were submitted to SignalP (http://www.cbs.dtu.dk/services/SignalP/) (Bendtsen et al. 2004) and WoLF PSORT websites (http://www.wolfpsort.seq.cbrc.jp/). However, the proteins that were predicted not to have a signal peptide were further analyzed using DGPI algorithm (http://www.129.194.185.165/dgpi/DGPI_demo_en.html) that can identify both signal peptide and GPI anchor addition sequences. Big-PI program (http://www.mendel.imp.univie.ac.at/gpi/plant_server.html) was also used to identify the GPI anchor addition sequences. Three-dimensional structure of the putative FLA domains was investigated using PHYRE (Protein Homology/analogY Recognition Engine) algorithm at http://www.sbg.bio.ic.ac.uk/~3dpssm. This web-based program searches first to find homologs to the query in public databases using the psi blast algorithm, then it searches Prosite and fold library before initiating loop modeling, cleft detection, and adding side chains. The three dimensional prediction is then visualized via Jmol software (http://www.jmol.org). The co-expression pattern of the Arabidopsis *FLA* genes was investigated using the Arabidopsis co-expression database hosted at the http://www.csbdb.mpimp-golm.mpg.de/csbdb/dbcor/ath.html website. This analysis may allow the identification of interacting molecules and develop hypothesis on the functional interactions of *FLA* genes.

Phylogenetic analysis

Fasciclin-like domains were identified in wheat and rice proteins using several programs as described above. The sequences of these putative fasciclin domains (~100 aa) were extracted and multiple alignments were carried out using ClustalW at http://www.ebi.ac.uk/clustalw or, for smaller datasets, T-COFFEE at http://www.ch.embnet.org. After manual editing, the alignments were displayed and shaded with GeneDoc program (http://www.psc.edu/biomed/genedoc). For phylogenetic analysis, another in-house java script was developed that combines PHYLIP 3.65 package (Felsenstein 1993), PHYML algorithm (http://www.atgc.lirmm.fr/phyml) (Guindon et al. 2005), and TreeView 1.6.6 (http://www.taxonomy.zoology.gla.ac.uk/rod/treeview.html), a program for displaying and manipulating trees. This java script allows for the creation of PHYLIP alignments that can be edited before submission for analysis under the JTT (Jones et al. 1992) or Dayhoff (Dayhoff et al. 1978) substitution frequency matrix with 1,000 bootstrap replicates as implemented in PHYML algorithm. The phylogenetic tree was then visualized with TreeView software.

Expression patterns

Various wheat tissues libraries were created from TIGR ESTs databases. The libraries were leaf (63743 ESTs); crown (12480 ESTs), roots (58148 ESTs), seeds (72462 ESTs), kernel (15253 ESTs), endosperm (11110 ESTs), anthers (9913 ESTs), and pistil (10161 ESTs). Electronic screening for *TaFLA* genes in these tissues libraries was performed using the blastN program and an EST with above 97% identity over at least 600 pb was counted as the exact match of the gene. The data is summarized in Table 1. Also, to confirm and complement this "digital northern" data, RNA-gel blot analysis was carried out on some *TaFLA* genes in leaf,

crown (the part of the seedling just above and below the ground from which the roots and shoots branch out). Briefly, the RNA was extracted using Tri Reagent according to the manufacturer's protocol (Molecular Research Center, Cincinnati). Full-length cDNAs were cloned by PCR from cDNA libraries and used as probes. Northern analyses were performed using standard procedures (Sambrook et al. 1989) and repeated at least three times using three different RNA preparations (biological replicate), and a typical result is presented in Figs. 4, 5, 6.

Results

Identification and cloning of putative wheat *FLA (TaFLAs)* genes

A java script was developed that performs several bioinformatics steps in a row and gives a final list of unique, putative wheat *FLA* sequences. As a first step, the java script performs a tBlastn (NCBI) search on the TIGR wheat EST databases (580,155 ESTs), using the 21 Arabidopsis FLA protein (AtFLA) sequences as queries. Arabidopsis FLA4 (At3g46550) gave two accession numbers NP_190239 and NP_915428, but the later number corresponds to a rice protein accession number. In a second step, the script collects all the hits ($E < 0.05$) obtained from each of the AtFLA proteins, which yielded a list of ~1017 EST sequences. Then the script removes redundancy (same accession number) from this first crude list, thus reducing the number to 143 ESTs. In the next step, the java script merges these ESTs into unique contigs using the CAP3 assembly program (Huang and Madan 1999) further reducing the number to 63 putative TaFLA sequences. In order to identify contigs that might correspond to different regions of the same gene, we performed several alignments using the clustalW program and manually compared the similarities with the closest full-length Arabidopsis FLA sequence. This step not only reduced the number of candidates to 43 sequences, but also allowed us to assess whether contigs were full-length or partial sequences. Short sequences (less than 300 pb) with low quality nucleotide sequences (~10–30% NNNs) were discarded which reduced the number to 38 putative TaFLA sequences. Four of these putative TaFLA sequences showed high similarity with maize genes (92–95% identity at the nucleic acid sequences), suggesting that they might be maize ESTs mistakenly deposited into wheat EST databases. Therefore, we decided not to include them in this study and to limit further characterization to 34 putative TaFLA contigs. Among these 34 wheat sequences 50% (17 contigs) were full-length clones. These 17 contigs were cloned and fully sequenced at least three times to avoid sequencing mistakes, and their identities were assessed by additional PCR using gene-specific primers. The full length clones that correspond to the remaining 17 incomplete sequences (boldface type in Table 1) were amplified by a PCR-based approach using gene-specific primers from various cDNA libraries prepared from enriched tissues (spike, developing endosperm, or stressed tissues). Each cloned gene was again fully sequenced at least three times to avoid sequencing mistakes. The 34 putative TaFLAs are listed in Table 1 along with the TIGR accession numbers. In addition, a total of 24 rice genes were identified and included in Table 1 along with their accession numbers.

The putative wheat and rice proteins have FLA protein characteristics

To confirm that the putative TaFLAs and OsFLAs belong to the AGP super-family, we first searched for characteristics that have been identified in AtFLA proteins such as: (1) the presence of at least one HRGP glycomodule (Nothnagel 1997; Bacic et al. 2000; Showalter 2001; Johnson et al. 2003; Tan et al. 2003), (2) the presence of a signal peptide at the N-terminus and in some cases an additional C-terminal signal for a glycosylphosphatidylinositol (GPI) anchor addition, and (3) the presence of at least one fasciclin-like domain. However, the classification of these molecules as HRGPs will require further experimental evidence such as their ability to bind β-glycosyl Yariv reagent or elucidating their pattern of *O*-glycosylation. According to the "Hyp contiguity hypothesis", usually a Hyp residue can be arabinogalactosylated if it is repeated in a noncontiguous way in motifs such as [Ala/Ser/Thr]-Hyp-X(0,10)-[Ala/Ser/Thr]-Hyp, where two consecutive Hyp are not separated by more than 11 amino acid residues (Schultz et al. 2002). However, Hyp repeats in [Ala/Ser/Thr]-Hyp-Hyp and [Ala/Ser/Thr]-Hyp-Hyp-Hyp glycomodules seem to be the attachment sites for linear arabinooligosaccharides (up to five residues) instead of AG polymers (Tan et al. 2003). To predict the occurrence of such glycomodules in our putative wheat and rice FLA proteins, we manually counted Hyp-containing sequence motifs in each of them and data are summarized in Table 1. Although, most of the putative TaFLAs and OsFLAs contain in their sequences these glycomodule motifs, only 5 OsFLAs and 5 TaFLAs have [Ala/Ser/Thr]-Hyp-Hp-Hyp motif (Table 1). It is noteworthy to point out that TaFLA30 and OsFLA22 (Os02g26290) seem to lack completely

Table 1 Putative fasciclin-like containing proteins identified in wheat and rice genomes

Name	Accession# (TIGR)	Size (aa)	Number of putative glycomodules				FLA-like domain	Signal peptide	GPI anchor
			[AST]-P	([AST]-P-X (0,10)-[AST]-P)$_x$	[AST]-P-P	[AST]-P-P-P			
Wheat proteins									
TaFLA1	TC256308	430	2	1	0	0	1	Yes	Yes
TaFLA2	TC255748	404	3	1	1	1	1	No	Yes
TaFLA3	TC235385	416	3	2	0	0	1	Yes	Yes
TaFLA4	TC240008	265	0	3	2	0	1	Yes	Yes
TaFLA5	TC237235	429	1	2	2	0	1	Yes	Yes
TaFLA6	TC235885	367	5	0	0	0	1	Yes	Short
TaFLA7	TC235887	342	4	0	0	0	1	Yes	Short
TaFLA8	TC235884	342	4	0	0	0	1	Yes	Short
TaFLA9	TC241226	264	0	2	1	0	1	Yes	Yes
TaFLA10	TC269370	265	0	2	2	0	1	Yes	Yes
TaFLA11	TC262593	255	0	2	0	0	1	Yes	Yes
TaFLA12	TC253009	276	0	2	2	0	1	Yes	Short
TaFLA13	BT009005	267	1	1	1	0	1	Yes	Yes
TaFLA14	TC234795	245	0	2	0	0	1	Yes	Yes
TaFLA15	TC249162	289	1	2	0	0	1	Yes	Yes
TaFLA16	CK216481	263	0	3	0	0	1	Yes	Yes
TaFLA17	CK208285	256	0	3	0	0	1	Yes	Yes
TaFLA18	TC255878	263	2	2	0	0	1	Yes	Yes
TaFLA19	CV761977	480	3	0	2	0	1	Yes	No
TaFLA20	TC276044	436	5	1	4	0	1	Yes	Yes
TaFLA21	TC244333	277	1	0	0	0	1	Yes	No
TaFLA22	TC245365	435	5	1	3	1	1	Yes	Yes
TaFLA23	TC259662	266	0	2	2	0	1	Yes	Yes
TaFLA24	BT008953	264	0	3	1	0	1	Yes	Yes
TaFLA25	TC266543	459	1	1	2	0	2	Yes	Short
TaFLA26	TC266544	460	1	1	1	0	2	Yes	Short
TaFLA27	CA705196	482	1	2	2	0	2	Yes	Short
TaFLA28	CK201849	402	1	2	0	0	2	Yes	Yes
TaFLA29	CK212728	310	5	1	0	0	1	Yes	Short
TaFLA30	TC236129	205	0	0	0	0	1	Yes	Short
TaFLA31	TC265873	298	1	2	1	1	1	Yes	Yes
TaAGP1	TC265874	294	2	2	1	1	0	Yes	Yes
TaFLA33	TC256781	265	1	1	0	0	1	Yes	Yes
TaFLA34	CA597113	298	0	2	0	1	1	Yes	Yes
Rice proteins									
OsFLA1	Os04g48490	431	1	1	2	0	1	Yes	Yes
OsFLA2	Os03g03600	401	0	2	1	1	1	Yes	Yes
OsFLA3	Os08g23180	415	2	3	1	0	1	Yes	Yes
OsFLA4	Os08g38270	271	0	2	2	1	1	Yes	Yes
OsFLA5	Os08g39270	274	0	2	0	0	1	Yes	Yes
OsFLA6	Os05g48900	272	0	2	0	0	1	Yes	Yes
OsFLA7	Os01g47780	266	1	1	0	0	1	Yes	Short
OsFLA8	Os01g06580	255	0	2	0	0	1	No	Yes
OsFLA9	Os05g07060	265	1	2	1	0	1	Yes	Yes
OsFLA10	Os09g30010	273	4	2	3	0	1	Yes	Yes
OsFLA11	Os09g07350	401	3	2	0	0	1	Yes	Yes
OsFLA12	Os01g62380	403	0	1	0	1	1	Yes	Short
OsFLA13	Os04g39600	263	0	2	0	0	1	Yes	Yes
OsFLA14	Os04g39590	277	0	1	1	2	1	Yes	Yes
OsFLA15	Os02g20560	268	1	2	0	0	1	Yes	Yes
OsFLA16	Os07g06680	479	1	1	2	0	2	Yes	Short
OsAGP1	NP1114086	151	0	0	1	0	0	Yes	No
OsAGP2	Os02g28130	148	1	0	2	0	0	Yes	Short
OsFLA19	Os02g20540	267	1	2	1	0	1	Yes	Yes
OsFLA20	Os02g26320	304	2	2	0	1	1	Yes	Yes
OsFLA21	Os02g49420	266	0	2	0	0	1	Yes	Yes
OsFLA22	Os02g26290	213	0	0	0	0	1	Yes	Short

Table 1 continued

Name	Accession# (TIGR)	Size (aa)	Number of putative glycomodules				FLA-like domain	Signal peptide	GPI anchor
			[AST]-P	([AST]-P-X (0,10)-[AST]-P)$_x$	[AST]-P-P	[AST]-P-P-P			
OsAGP3	Os04g21570	278	0	2	1	0	0	Yes	Yes
OsFLA24	Os03g57460	474	0	2	2	0	2	Yes	No

The number of fasciclin-like domains is indicated for each protein. The TIGR gene accession number and a summary of the computer-based predictions of the secretion signal peptide and the GPI-anchoring signals are given for each protein. The number of the putative glycomodules present in each protein sequence was counted manually. Boldfaced FLAs are sequences that were completed using PCR-based methods

such AGP glycomodules, and TaFLA6, 7, 8, and 21 have only scattered [Ala/Ser/Thr]-Pro di-peptide sequence separated by more than ten amino acids. Two additional characteristics are observed in known FLA proteins: they are usually secreted proteins and may be attached to the plasma membrane via a GPI anchor. Therefore, they should possess an N-terminus signal peptide to target them to the secretory pathway and may have a C-terminus signal sequence recognized by a transamidase that replaces this peptide sequence with a GPI module. Putative TaFLA and OsFLA proteins were thus analyzed for the presence of these signal sequences using several prediction programs such SignalP (Bendtsen et al. 2004), WoLF PSORT (Horton et al. 2006), big-PI (Eisenhaber et al. 2003), and DGPI (Kronegg and Buloz 1999). As expected, all the putative wheat and rice FLA proteins, except TaFLA2 and OsFLA8, were predicted to be secreted by at least one of these algorithms (Table 1). Similarly, when putative AtFLA proteins were run on these programs (as control), they were predicted to be secreted proteins except for AtFLA20 protein (At5g40940). Further experiments will be required to confirm these predictions and optimize the current available algorithms. A putative FLA protein is considered GPI-anchored protein if at least one of the two programs (Big-PI and DGPI) predicts the presence of the motif at its C-terminus region. Twenty-three putative TaFLAs and 18 putative OsFLAs were predicted to have a GPI anchor sequence signal by at least one program. However, TaFLA19, OsFLA17 (NP1114086), and OsFLA24 (Os03g57460) were predicted not to be GPI-anchored proteins by both programs (Table 1). On the other hand many putative TaFLA proteins (11 proteins) have a short C-terminus hydrophobic region, which prevented GPI anchor signal prediction analysis, suggesting that the number of GPI-anchored TaFLAs may be underestimated. Therefore, further experimental evidence will be needed to determine the anchoring position sites, which will help in optimizing the prediction programs. When the CDD program at NCBI was used to search the conserved domains database using as queries the putative TaFLA and OsFLA proteins, one fasciclin-like domain was identified in 24 wheat and 19 rice proteins, and two FLA domains were found in four wheat (TaFLA25, 26, 27, and 28) and in two rice (Os07g06680 and Os03g57460; Table 1) proteins. However, because fasciclin domain sequences are not well conserved (Kawamoto et al. 1998), the CDD program could not identify such domain in six of the wheat proteins (TaFLA29, 30, 31, 32, 33, and 34). Similarly, this program could not identify regions with homology to fasciclin domain in several AtFLA proteins used as control (data not shown). Thus, we tried other prediction algorithms such as "Motif Scan" or "Pfam", which confirmed that both TaFLA31 and TaFLA34 have indeed one region with sequence similarity to fasciclin domain, as defined by the consensus sequence "smart00554" in public database. However, only a region with weak amino acid sequence similarity was identified in TaFLA29, 30 and 33. All prediction programs failed to identify such similarity in TaFLA32, OsFLA17, 18, and 23 (Table 1). These last four putative FLAs could simply be AGPs, which is not surprising because we have used full-length AtFLA protein sequences to search the databases. Hence, we expected to find among the hits sequences with fasciclin-like domains and/or AGP glycomodules. The PAST contents of these four proteins are higher than 40%, except for OsFLA17 (29%), but the later contains an A-P-P motif characteristic of AGPs. Thus, we concluded that these four sequences are AGPs and are indicated in Table 1 as TaAGP1, OsAGP1, 2, and 3 respectively. The putative fasciclin domains identified in TaFLA, AtFLA, and OsFLA proteins were aligned using the clustalW program and the alignment shows the conserved regions depicted in the smart00554 motif sequence, namely, the H1 conserved region is characterized by the following sequence [Ser/Thr]-[Val/Leu/Ile]-Phe-Ala-Pro-X-[Asp/

Glu/Asn]-X-Ala, and H2 is characterized by [Val/Leu/Ile]-[Phe/Tyr/His/Gln]-X-[Val/Leu/Ile]-X-X-[Val/Leu/Ile]-[Val/Leu/Ile]-[Val/Leu/Ile]-Pro sequence, where X can be any amino acid (Fig. 1). The third conserved region of the domains is located between H1 and H2 regions and is characterized by a two-residue motif, Tyr-His, which is usually flanked by [Leu/Val/Ile]-[Leu/Val/Ile] residues (Fig. 1). These residues have been shown to play roles in integrin binding in animal cells (Kawamoto et al. 1998; Kim et al. 2002). Although fasciclin-like domains were predicted in AtFLA, TaFLA, and OsFLA sequences, their alignment shows that all the regions of the domain are not conserved in some of these proteins (Fig. 1). Figure 1 also shows that Thr residues in the H1 region are completely conserved in the fasciclin super family. Similarly, an Asn, Glu or Asp residues (charged amino acids), always occupies the sixth position after the Thr residues, except for the AtFLA17.2 domain where Ala residue is found at this position (Fig. 1). The C-terminus section of H2 regions is rich in small hydrophobic amino acids such as Val, Leu and/or Ile in virtually all known fasciclin domains (Coult et al. 2003; Shi et al. 2003; Johnson et al. 2003; Lafarguette et al. 2004). Figure 1 shows that plant H2 regions have also conserved this characteristic. However, His residue usually conserved in all known animal fasciclin domains, is mostly substituted in plant proteins by another ring-containing amino acid (Phe or Tyr), or by another amide-containing amino acid (Gln) (Fig. 1). Furthermore, the alignment shows that the section between H1 and H2 motifs of known fasciclin domains is also conserved in plant proteins. However, the His residue conserved in this region is missing in some proteins such as AtFLA20.1, AtFLA8.1, AtFLA10.1, AtFLA1.2, OsFLA6, TaFLA15, and TaFLA1. Finally, it is noteworthy to mention that like in Arabidopsis FLA proteins, few wheat proteins have a peptide stretch (~6 amino acids) that is rich in either charged (i.e., Lys, Asp, Glu) or hydrophobic (i.e., Val, Leu, Ile) amino acids near the C-terminus. For example, TaFLA3, 15, 28, and 33 have a combination of poly-Lys and poly-Asp stretches near the C-terminus, in addition to two AGP domains (Fig. 2). The C-terminus regions of TaFLA16 and 17 have stretch with both Lys and Asp residues mixed together in addition to the AGP domain (Fig. 2). However, TaFLA21 has instead a hydrophobic region rich in Leu, Val, and Ile residues flanking the FLA domain. This TaFLA is lacking the AGP domain and GPI-anchoring signal (Fig. 2). These charged or hydrophobic regions may play a role in stabilizing/orienting the protein on the cell surface.

Molecular modeling of plant fasciclin-like domains

To get insight about the structure/function relationships in plant fasciclin-like domains, fold recognition methods were used to predict whether the plant FLA domain protein sequences have folding similarity to the recently resolved FasI domain 4 crystals (Coult et al. 2003). Fold recognition uses threading process to align the query protein sequence with the known crystal structure (Godzik 2003). PHYRE is a web-based algorithm that allows performing such alignment and gives the secondary structure predictions (α-helices and β-strands), with the 3-D coordinates, the E-value, the estimated precision, and sequence identity percentile. This structural analysis will also give additional support to the assignment of these putative proteins as members of the FLA super-family and may provide a rational basis to develop educated experiments for the evaluation of the role of certain protein regions in adhesion and signaling mechanisms. To test the validity of the sequence-to-structure alignment algorithm, we used as a control the smart00554 sequence. As expected the threading analysis predicted that smart00554 sequence should adopt FasI domain 4 fold with an estimated precision of 100% (Fig. 3b). Smart00554 fold is organized in seven beta sheets (β1-7) covering the two sides (back to back) of the domain. One side is covered by β1-β2-β7-(β6 end) sheets and the back side by the anti-parallel β4-β5-(start of β6) sheets. Both series form a compact triangular shape with the β6 sheet split over both sides. The sharp edge of the triangle is formed by the α3 and α4 helices and β3 strand forms the base of the open end of this triangle (Fig. 3b). The model indicates also that the fold is maintained by hydrogen bounds between small hydrophobic amino acids distributed along the β-strands. When the program was run with the 92 plant fasciclin-like domains, 86% of the FLA domains were predicted to adopt folding pattern to the FasI domain 4 with an estimated precision above 70% based on amino acid sequence identity between 20 and 95%. It is noteworthy to indicate that these identity scores are lowered because the FasI domain 4 sequence used for similarity is about 30% longer than the plant sequences. Examples of plant fasciclin-like domain folds are included in Fig. 3. Although, the estimated precision was below 50% for 5 Arabidopsis (AtFLA10.1, AtFLA17.1, AtFLA19, AtFLA20.1, and AtFLA20.2), and one wheat (TaFLA34) fasciclin-like domains, the PHYRE algorithm could still build folds for these domains using FasI crystal structure.

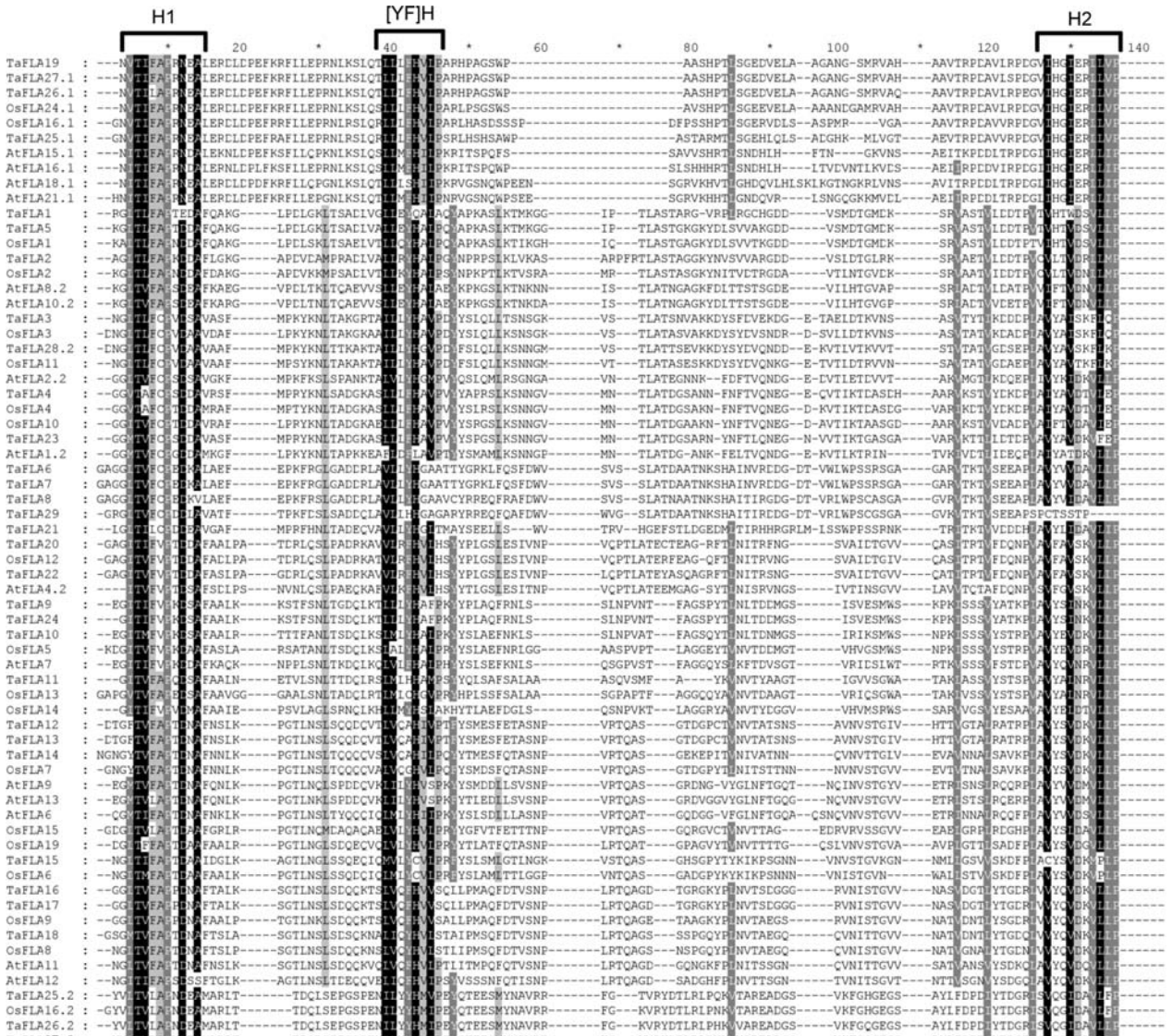

Fig. 1 Multiple sequences alignment of the fasciclin-like domain from Arabidopsis, rice and wheat. The alignment was generated by ClustalW program and manually edited. Residues in positions conserved at 90, 75, and 55% are shaded in *black*, *dark gray*, and *light gray*, respectively. *Dashes* represent gaps introduced by the program for optimal alignment. The three conserved regions characteristic of fasciclin domains (H1, H2, and [FY]H) are indicated on the *top* of the alignment. In proteins with more than one FLA domain have the annotation "*.1*" or "*.2*" to indicate the position of the domain in the protein starting from the N-terminus end of the protein

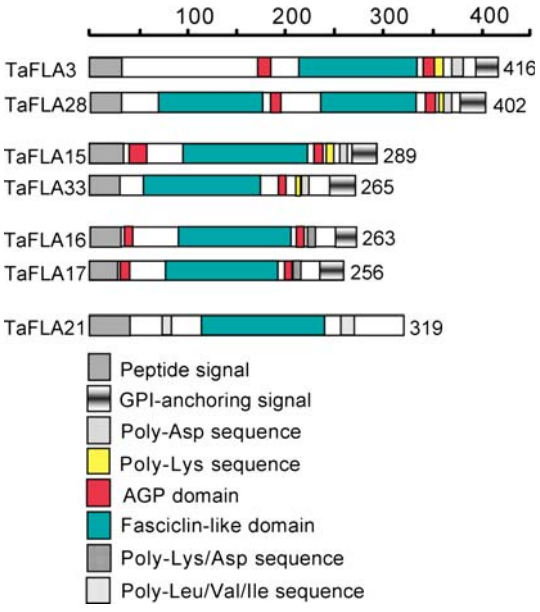

Fig. 2 Representation and location of various domains and sequence stretches found in some wheat FLA proteins. The regions indicated are fasciclin domains, AGP domain, signal peptide, GPI signal peptide, and the stretches of poly-Lys, poly-Asp, and poly-Leu/Val/Ile sequences

Putative *TaFLA* genes show differential expression patterns in various tissues and under environmental stresses

Since no microarray data is available for wheat, we used an alternative strategy that consists of downloading all TIGR wheat EST databases (580,155 ESTs) and then grouping them into EST libraries of various tissues. These libraries were then searched for representatives of each of our *TaFLA* sequences using the blastN program. ESTs with above 97% identity over at least 600 pb were counted as the exact match of the *TaFLA* sequence. Table 2 summarizes these results along with the total number of ESTs in each library. According to this "digital northern" data all *TaFLAs* are mostly expressed in seeds and roots. Only 6 of the 34 *TaFLA* genes (*TaFLA3, 6, 15, 25, 26, and 31*) are highly expressed as judged by the total number of their ESTs found. The data show also that some of the genes are tissue specific. For example, *TaFLA5* and *31*genes for which the most ESTs identified come from endosperm cDNA libraries. *TaFLA14, 15, 25*, and *26* have ESTs found mostly in roots, and *TaFLA30* and *34* in anthers. *AtFLA33* was the only gene with one EST present in pistil cDNA libraries. *TaFLA22* gene, the closest FLA to *AtFLA4.2* (*SOS5* gene) was expressed at a low level in seeds (two ESTs). In order to confirm and complement this expression data, we carried out northern blot analysis on some *TaFLA* genes in leaf, crown, and roots. Our experimental data confirmed the expression pattern of *TaFLA2, 3, 4, 9, and 12* transcripts, namely *TaFLA2* and *9* seem to be expressed preferentially in roots, *TaFLA4* in crown, and *TaFLA12* in root and leaf (Fig. 4). The only discrepancy is that *TaFLA14*, which has several ESTs in roots and leaves in TIGR databases, seems to be weakly

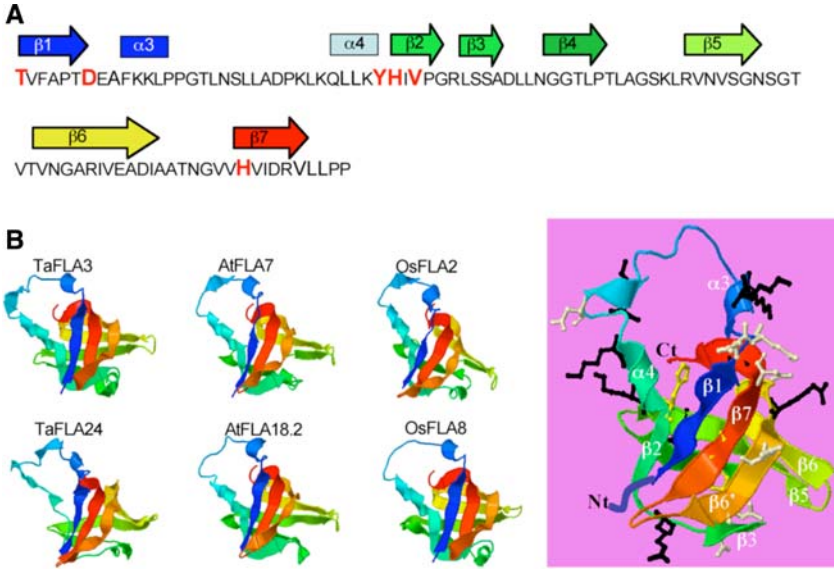

Fig. 3 Three-dimensional folding model of the smart00554 motif along with typical predicted folding patterns of some plant fasciclin-like domains. **a** smart00554 amino acid sequence with the conserved amino acid residues highlighted in *bigger letters*, and its putative secondary structures (above the amino acid sequence). **b** Cartoon representations of smart00554 and six typical examples of plant FLA domains (Arabidopsis, wheat, and rice) fold model as predicted by PHYRE program using as model fasciclin I domain 4 of βig-h3. Positively and negatively charged residues are drawn in *black* and *white*, respectively in the cartoon of smart00554

expressed according to our experimental data (Fig. 4, Table 2). We further tested *TaFLA16*, *18, 19* and *29* transcripts for which no clear conclusion could be drawn from the ESTs distribution (Table 2). Our northern blot analysis indicated low expression of *TaFLA16, 19*, and *29* transcripts in the three tissues; however, the *TaFLA18* gene was detected mostly in leaves and roots (Fig. 4, Table 2). The expression patterns of several *TaFLA* genes were determined in wheat shoots in response to abiotic stresses such as cold, heat, salt, and dehydration in addition to ABA treatments. Our data showed that most of these genes were not affected (data not shown). This is not surprising because most of the *TaFLA* genes were expressed in seeds or roots and only few were detectable in above ground tissues (shoots) (Fig. 4, Table 2). However, the transcript of any up-regulated gene would be seen easily by northern gel blot analysis. For example during cold treatment, *TaFLA12* and *TaFLA9* transcripts showed a maximum accumulation in roots of plants that were acclimated for 6 days at 4°C, but the transcripts returned to normal levels after a longer period of acclimation (36 days) (Fig. 5). A similar pattern was observed also for *TaFLA18* (data not shown). A slight increase in *TaFLA14* transcript was also observed in roots after 1 and 6 days acclimation (Fig. 5). However, regarding other stresses, the transcript levels of *TaFLA12* and *9* did not change in above ground tissues in response to heat or salt treatment, but were down regulated by ABA treatment and dehydration treatments (Fig. 6). The *TaFLA3* transcript was down regulated in all the treatments except heat-induced stress. Interestingly, *TaFLA14* was the only gene that was up regulated in response to 70% dehydration-induced stress (Fig. 6). *TaFLA20* and *22*, homologs to the *SOS5* gene in Arabidopsis, were weakly expressed and the transcripts could not be detected in leaf, crown, or root tissues even after 16 h exposure.

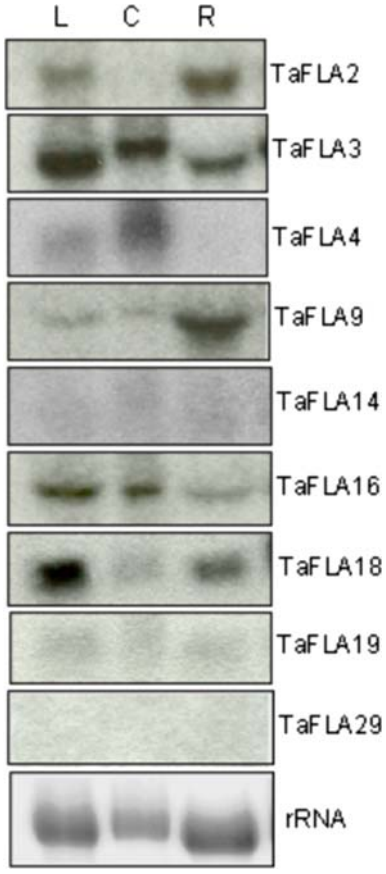

Fig. 4 Expression of *TaFLA* transcripts in leaf (*L*), crown (*C*), and roots (*R*) of 7-day-old wheat seedlings using RNA gel blot analysis. In all lanes, 10 μg total RNA was used and the 28S ribosomal RNA stained with ethidium bromide is included as loading controls for each lane. Membrane exposure was 16–48 h depending on the gene

Phylogenetic analysis of plant fasciclin-like domains

Phylogenetic analysis could be a useful way to predict the physiological function of genes. It would be interesting to know the evolutionary history of plant fasciclin-like domains. For example, are they the result of gene duplication? did the duplicated domains evolved in similar way? Would it be possible to connect the evolutionary clustering to a functional clustering? To be able to make conclusive phylogenetic analysis, we included FLA domains from four species representing dicots (Arabidopsis), monocots (wheat and rice), and gymnosperms (*Pinus taeda*). The pine FLAs domains were collected from EST public databases (TIGR). The advantage of increasing species sampling is to increase the possibility of deducing the point at which gene duplications occurred. In addition, we used only FLA domains because the repetitive nature of AGP glycomodule found in FLA-AGP proteins can be problematic in making a reliable alignment of FLA proteins, which is essential in a phylogenetic analysis such this one. The term orthologs or homologs are used here to describe putative functionally equivalent domains in Arabidopsis, wheat, rice, and pine based entirely on sequence similarity and clustering. Our phylogenetic analysis indicates that plant fasciclin-like domains could be grouped into at least eight groups (I–VIII, Fig. 7). It seems that group-I and the other groups (II–VIII) are the result of an ancient duplication (Fig. 7). Although bootstrap support values are low for the main nodes, it is still possible to distinguish the clustering (groups-I–group-VIII). Interestingly, these groups seem to evolve differently. For example, group-I includes FLA domains that are less divergent

Table 2 Wheat FLA ESTs found in publicly available in TIGR cDNA libraries

Name	Total ESTs	Leaf 63743	Crown 12480	Root 58148	Seed 72462	Endos 11110	Kernel 15253	Anthers 9913	Pistil 10161
TaFLA1	5	–	–	–	1	1	2	–	–
TaFLA2	7	–	1	3	3	–	–	–	–
TaFLA3	37	8	9	–	12	–	2	1	–
TaFLA4	5	–	1	–	2	–	1	–	–
TaFLA5	7	–	–	–	–	4	1	–	–
TaFLA6	21	–	–	–	15	–	2	–	–
TaFLA7	4	–	–	–	2	–	–	–	–
TaFLA8	8	–	–	–	8	–	–	–	–
TaFLA9	5	–	1	3	–	–	–	–	–
TaFLA10	10	1	1	–	3	–	1	–	–
TaFLA11	2	–	–	1	–	–	–	–	–
TaFLA12	5	1	–	1	1	–	2	–	–
TaFLA13	1	*	*	*	*	*	*	*	*
TaFLA14	17	3	–	8	–	–	–	1	–
TaFLA15	27	2	1	8	1	–	–	–	–
TaFLA16	1	*	*	*	*	*	*	*	*
TaFLA17	1	–	1	–	–	–	–	–	–
TaFLA18	6	*	*	*	*	*	*	*	*
TaFLA19	1	*	*	*	*	*	*	*	*
TaFLA20	2	*	*	*	*	*	*	*	*
TaFLA21	2	*	*	*	*	*	*	*	*
TaFLA22	2	–	–	–	2	–	–	–	–
TaFLA23	2	–	–	–	1	–	–	–	–
TaFLA24	1	*	*	*	*	*	*	*	*
TaFLA25	15	–	1	7	1	–	–	2	–
TaFLA26	13	1	1	7	2	–	–	–	–
TaFLA27	1	–	–	–	–	–	1	–	–
TaFLA28	1	–	–	1	–	–	–	–	–
TaFLA29	1	*	*	*	*	*	*	*	*
TaFLA30	7	–	–	–	–	–	–	3	–
TaFLA31	23	–	–	–	1	15	4	–	–
TaAGP1	12	–	–	–	9	3	–	–	–
TaFLA33	4	–	–	–	3	–	–	–	1
TaFLA34	1	–	–	–	–	–	–	1	–

The libraries were generated from TIGR cDNA libraries of various tissues: *L* leaf (63743 ESTs), *C* crown (12480 ESTs), *R* roots (58148 ESTs), *S* seed (72462 ESTs), *K* kernel (15253 ESTs), *Endos* endosperm (11110 ESTs), *A* anthers (9913 ESTs), and *P* pistil (10161 ESTs). The ESTs with above 97% identity over at least 600 bp were counted as an exact match of the gene. The "*" indicates the lack of conclusive information for these transcripts

and have the highest amino acid sequence similarity with "smart00554" domain. On the other hand, group-II–group-VIII showed relatively more diversified monophyletic clusters and seem to be the result of more recent duplications. The most recent duplications yielded the two FLA domains of AtFLA17 and 20 (group-III) and PtFLA13 domains (group-IV). Interestingly, the two fasciclin-like domains of SOS5 protein (AtFLA4) involved in salt sensitivity of Arabidopsis (Shi et al. 2003) seem to be also a result of duplication but the two copies (in group-V and group-VII) have evolved differently. Both of SOS5 FLA domains have homologs in other species. Another interesting observation from this phylogeny analysis is that, although most of the FLA domain groups have representatives in dicots, monocots, and gymnosperms, some groups seem to be confined to certain species. For example, group-III composed of AtFLA17.1, AtFLA17.2, AtFLA20.1, AtFLA20.2 and AtFLA19 are found only in Arabidopsis, group-IV (PtFLA13.1, PtFLA13.2, and PtAFLA9) only in pine, and group-VIII (TaFLA6-8, and TaFLA21, 29) in wheat (Fig. 7).

Discussion

In this work, we describe the features of several FLA proteins from wheat and rice in an effort to elucidate their importance in cereal grain development. We have developed an in-house java script to screen wheat ESTs public collection in search for homologs to AtFLA proteins. This search indicates that wheat has at least 33 putative *FLAs* genes, a number that is higher than in Arabidopsis genome (12 more genes).

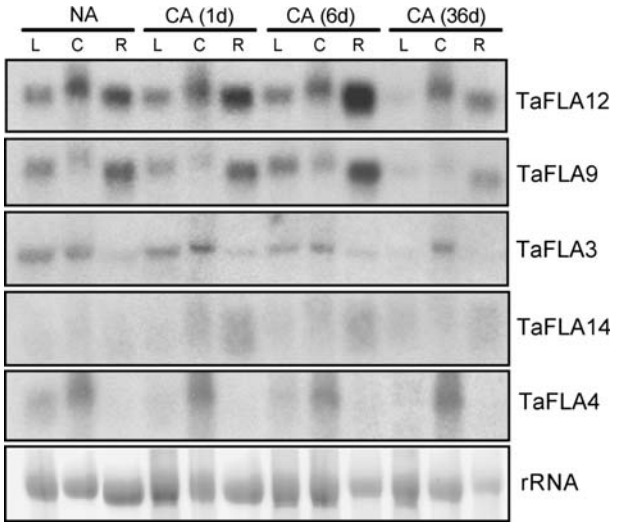

Fig. 5 Expression of *TaFLA* transcripts in leaf (*L*), crown (*C*), and roots (*R*) of wheat seedlings that were cold acclimated for 1, 6, and 36 days at 4°C as determined by RNA gel blot analysis. Northern conditions were similar to those indicated in Fig. 4. *NA* stands for nonacclimated plants grown for 7 days; *CA* (*1d*), *CA* (*6d*), and *CA* (*36d*) stand for plants that were cold acclimated for 1, 6, and 36 days

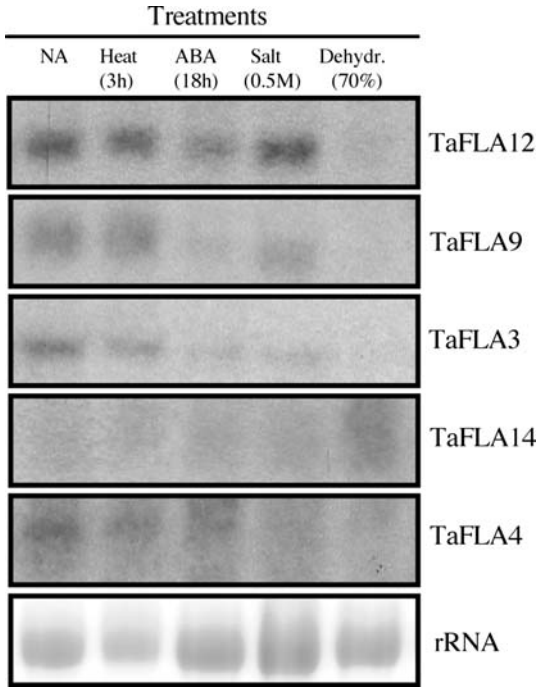

Fig. 6 Effects of various treatments on the expression of *TaFLA* transcripts in wheat seedling shoots. Northern blot conditions are as indicated in Fig. 4. *NA* nonacclimated seedlings grown for 6 days; salt, plants treated with 500 mM NaCl for 18 h, *ABA* plants treated with 0.1 mM ABA for 18 h, *dehydr.* plants exposed to 70% (w/v) of polyethylene glycol (MW 8 kD)

However, because wheat is a hexaploid plant having three copies of the same genome (ABD), some of the identified putative *TaFLAs* might be copies of the same gene. Usually, copies of the same gene would share high nucleic acid sequence identity (>97%) in the coding regions and low similarity at the 3′ and 5′UTR regions. Indeed, sequence comparisons indicated that the following pairs: TaFLA6 and 7; TaFLA9 and 24; TaFLA12 and 13; TaFLA16 and 17; TaFLA20 and 22; and TaFLA26 and 27, have coding regions that are 98–100% identical at the nucleic acid level, and the 3′ and 5′UTR regions have low similarity (<30%), which strongly supports the conclusion that each sequences of these pairs are copies of the same gene. Thus, the number of unique putative *TaFLA* genes can be reduced to 28. However, despite the presence of large number of wheat EST in the public databases, it is possible that we did not identify all *TaFLA* genes. Also, since many of the sequences are from EST datasets, the accuracy of the checking, even with the stringent criteria used, means that the prediction of some of the motifs and/or signal may be inaccurate. The deduced amino acid sequences of wheat and rice *FLA* cDNAs gave proteins with sizes ranging between 150 and 480 amino acids. Computer-based analyses showed two main differences between the cereal proteins and Arabidopsis FLAs. First, a low number of FLAs predicted to have two fasciclin-like domains (4 TaFLAs and 2 OsFLAs) in comparison to Arabidopsis (10AtFLAs). Second, while all AtFLAs have *O*-glycosylation sites (AGP glycomodules), 5 TaFLAs and 1 OsFLAs may potentially lack these *O*-glycosylation sites. Indeed, the Pro residues in these proteins do not comply with "Hyp contiguity hypothesis" as described by Tan et al. (2003) (Table 1). For example, TaFLA30 and OsFLA22 (Os02g26290) lack completely the AGP glycomodules. This observation is new because all putative plant FLA proteins identified thus far have the fasciclin-like domain associated with an AGP or EXT glycomodules (Shi et al. 2003; Johnson et al. 2003; Lafarguette et al. 2004). In addition, four wheat proteins (TaFLA6, 7, 8, and 21) do not have well-defined AGP glycomodules, instead they have only scattered [Ala/Ser/Thr]-Pro di-peptide sequence separated by more than ten amino acids. If the fasciclin domains of these four wheat FLA proteins were not included in the estimation of the Pro, Ala, Ser, and Thr residues (PAST) contents, the values observed were between 25 and 35% (depending of the region of the proteins), which are still below the 50% content characteristic of AGPs. It is possible that these scattered Pro residues may not undergo post-translational *O*-glycosylation, however it has been shown that sporamin, a non-AGP protein, undergoes arabinogalactosylation (as in AGPs)

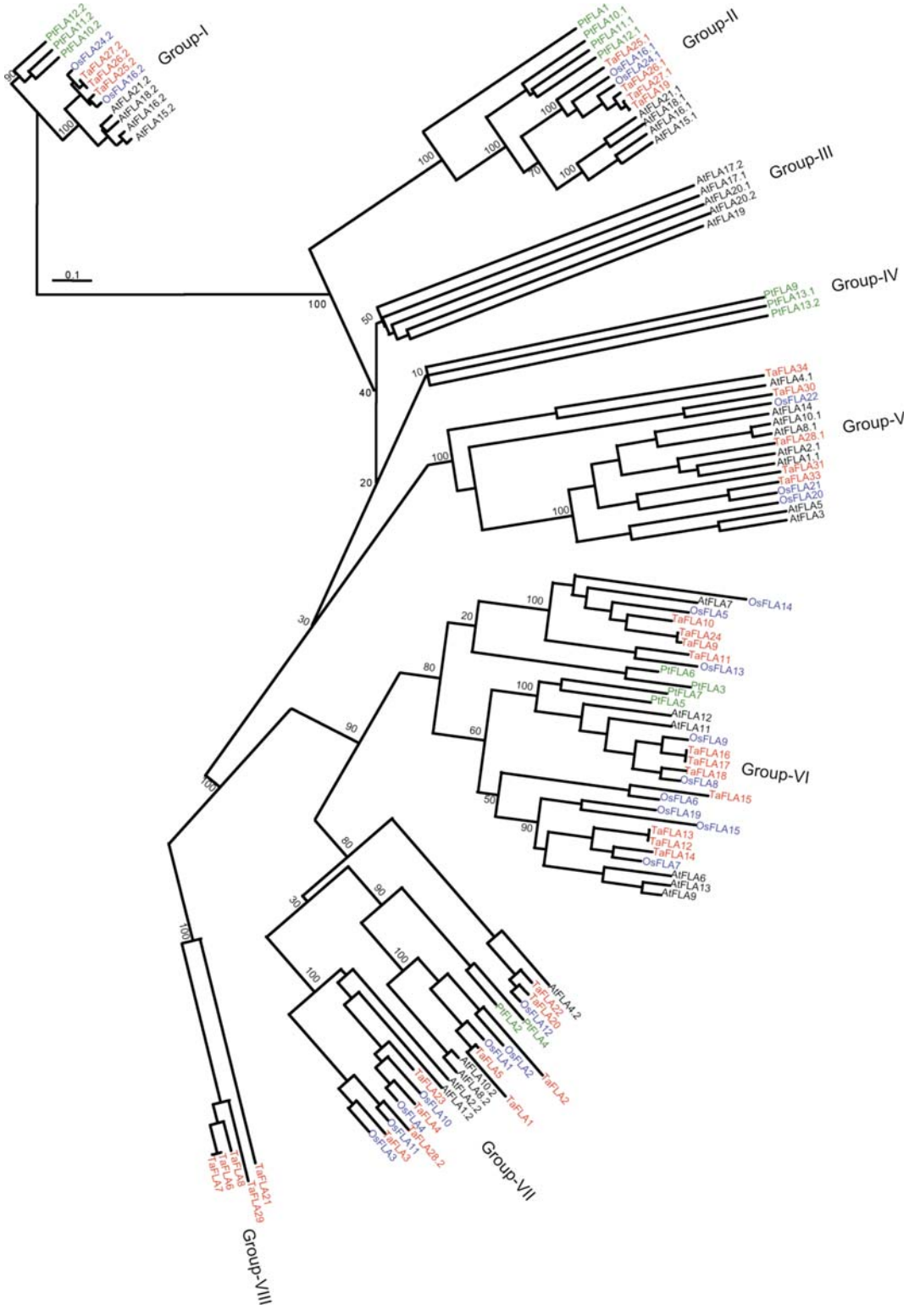

Fig. 7 Phylogenetic tree representation of the putative fasciclin domains (about 100 amino acids), from Arabidopsis (AtFLAs), wheat (TaFLAs), rice (OsFLA), and pine (PtFLAs). Protein sequences were aligned with ClustalW and manually edited, and a tree was constructed under the JTT substitution model using the PHYML program. Bootstrap support values of 1,000 trials are indicated

when expressed in tobacco BY2 cells (Matsuoka et al. 1995). Sporamin protein has a unique Pro residue in its sequence located at position 36 and the amino acids surrounding this Pro36 seem to be critical for its efficient hydroxylation and arabinogalactosylation (Shimizu et al. 2005). Thus, it would be of interest to investigate Pro *O*-glycosylation in TaFLA6, 7, 8, and 21. In a recent work (Shimizu et al. 2005), a more general rule for Pro hydroxylation and *O*-glycosylation was proposed that includes non-AGP proteins. In this rule, the motif required for efficient Pro hydroxylation and *O*-glycosylation consists of [AVSTG]-P-[AVST]-[GAVPSTC]-[APS or acidic], however this rule still does not apply to the Pro residues in TaFLA6, 7, 8, and 21 proteins because they are part of a different motif, namely A-P-[EQ]-[EL]-[EPR]. Therefore, a more general Hyp glycosylation rule is still needed. Although several AGPs have been characterized from monocots such as lolium (Gleeson et al. 1989), maize (Kieliszewski et al. 1990), and wheat (Fincher et al. 1974), details are still lacking regarding the composition and structural variations of the glycan portions. It would be interesting to express some *TaFLA* genes in tobacco cells and investigate the glycosylation pattern of the proteins. This will help extend the general rule of Hyp *O*-glycosylation to include non-AGP (sporamin) and the new types of AGP proteins (TaFLA6, 7, 8, and 21). The physiological significance of these differences is still unknown. Beside these differences, wheat and rice FLAs share many other common features with Arabidopsis proteins. For example, they all are predicted to be secreted proteins except for TaFLA2 and OsFLA8 for which no algorithm could identify a signal peptide in their sequences. Only TaFLA19, OsFLA17 and 24 were not predicted to be GPI-anchored to the plasma membrane. However, about 25% have shorter C-terminus amino acid sequences to allow a correct prediction suggesting the need of improvement of the current GPI prediction algorithm and experimental confirmation to support these predictions. Taking into consideration the current prediction, the number of TaFLA and OsFLA proteins predicted to have GPI anchor sequence (70%) is comparable to Arabidopsis proteins. RNA gel blot analyses as well as a database EST survey indicate that most of the *TaFLA* genes are expressed in reproductive tissues (i.e., seeds) similar to Arabidopsis *FLA* genes. This finding strongly suggests that these proteins are important in plant development where an exact coordination of embryogenesis steps is required. Indeed, FLA can play a key role in cell-cell and cell-environment communication by establishing physical contact with neighboring cells or with their extracellular matrices (cell walls). This physical connection could be a central piece in signal transduction by which the surrounding environment (e.g., neighboring cells, external variation etc.) would be sensed by the cell. The physical interactions may involve a fasciclin-like domain, which raises the question of how these plant domains mediate such an interaction and what the mechanism might be. In order to understand the mechanism of such interaction, the FAS1 domain of the insect cell was used as a model to create 3-D folds of putative plant fasciclin domains using the PHYRE algorithm. With the exception of 6 fasciclin-like domains (AtFLA10.1, AtFLA17.1, AtFLA19, AtFLA20.1, AtFLA20.2, and TaFLA34), the other 86 plant fasciclin-like domains were predicted to have folding homology to FasI domain 4 with estimated precision above 50%. These six domains may not be functional or the prediction program may need to be trained on more possible folds, which may require making crystals from the putative plant fasciclin-like domains. However, several observations can be drawn from this preliminary structural analysis: (a) hydrophobic amino acids (i.e., Ala, Ile, Leu, Val, Pro, Phe) are spread along the β-sheets and are responsible of maintaining the 3-D structure of the domains through hydrogen bonds between them. Phe and Pro residues are oriented toward the central pocket of the model. The small nonpolar amino acids are generally nonreactive and may not be involved in interactions, and (b) All domains have the negatively charged (i.e., Asp, Glu) and Tyr residues concentrated along the β1-α3-α4-β2 edge, while the positively charged amino acids (i.e., Lys, His, Arg) are concentrated on the back side of the folds. This distribution of highly charged surfaces could provide an excellent way of mediating protein–protein interactions via electrostatic forces as found in many other adhesion molecules such as CD2 and CD58, where charge instead of shape complementarity seems to play an important role in recognition (Wang et al. 1999). It has been shown that CD2 and domain 1 in FasIII have a striking structural similarity, which may suggest a similar interaction mechanism. In this case the interacting surfaces are hydrophilic and rich in charged residues such as Trp, Tyr, Arg and Lys. Our understanding of the molecular basis of interactions mediated by plant fasciclin-like domains is limited, however, extrapolation of the work carried out on the animal fasciclin and immunoglobulin super-family (Boyington et al. 2000; Wang et al. 1999; Soroka et al. 2002) to plant fasciclin domain model may allow predictions to be made concerning regions within the fold that mediate binding through homophilic (between similar fasciclin-like domains) or heterophilic (different

fasciclin-like domain or other interacting proteins) mechanism. Several *AtFLA* genes are found to be expressed at the same time and in the same tissues (according to the "*A. thaliana* Co-Response Database"), which may suggest the possibility of interaction between these FLA proteins via homophilic or heterophilic mechanisms. For example, *AtFLA18* gene is positively co-expressed with *AtFLA2*, *9*, and *13* genes suggesting they may interact with one another. Similarly, *AtFLA1* is co-expressed with *AtFLA7*, *8*, and *10*; and *AtFLA7* transcript is co-expressed with *AtFLA1, 2, 8, 9, 10, 11, 12*, and *16*. Phylogenetic analysis seems to indicate that plant FLA domains are the result of several duplications and the duplicated copies evolved at different paces. This analysis suggests also that the most recent duplications occurred in group-III and group-IV because the two domains of the proteins still clustering together. In these two groups Arabidopsis FLA19 and Pine FLA9 did not have FLA domain duplication. Interestingly, it seems that some FLA proteins lost one of the domains. For example, wheat FLA19 and Pine FLA1 both in group-II have only one fasciclin-like domain but clustered with other FLA proteins having two FLA domains (group-II). In term of functional prediction, it is difficult to make clear conclusions because of the lack of experimental data on most of clustered FLA proteins. The only clear example on the importance of FLA domain in plants is the salt overly sensitive Arabidopsis mutant (*sos5*). The mutation (Ser to Phe) in the most conserved region of FLA4.2 domain produced a mutant with defects in root cell wall structure and growth that makes, in turn the plant hypersensitive to high salt concentrations (Shi et al. 2003). AtFLA4.2 domain in SOS5 protein clusters with wheat FLA20 and 22, and rice FLA12 domains (group-VII). All of these FLA domains have a Ser residue at the same position suggesting similar physiological functions. It would be interesting to confirm this prediction with rice mutant. In addition, RNA blots analysis showed that *TaFLA12* and *TaFLA9* are specifically up regulated by cold treatment in roots and *TaFLA14* is up regulated by dehydration stress. The three FLA proteins belong to group VI, the closest to group VII. Taking these data together, we are tempted to conclude that group VI and VII may function in signaling pathway during abiotic stresses such salt, dehydration, and cold. Again, it would be of interest to evaluate whether AtFLA6, 7, 9, and 13, the Arabidopsis homologs of TaFLA9, 12, and 14 are also regulated by cold and/or dehydration treatments. It worth mentioning that group VII includes also AtFLA11, a protein that was shown to be involved in secondary wall formation (Brown et al. 2005; Persson et al. 2005). In conclusion, the characterization of wheat and rice FLAs and the structural analysis described above may establish the foundation for more detailed analysis, which could help design experiments where mutations in fasciclin-like domains may shed the light on the function of these proteins and identify the precise region of the domains that are important in the interaction.

Acknowledgments Author would like to thank Dr. Allan Showalter for critical reading of the manuscript and for valuable comments, and Matthew Shipp for his assistance in making the Figures. Dr. Harvey Ballard is gratefully thanked for his precious discussion on FLAs phylogeny.

References

Bacic A, Currie G, Gilson P, Mau S-L, Oxley D, Schultz C, Sommer-Knudsen J, Clarke AE (2000) Structural classes of arabinogalactan- proteins. In: Nothnagel EA, Bacic A, Clarke AE (eds) Cell and developmental biology of arabinogalactan-proteins. Kluwer Academic/Plenum Publishers, Dordrecht, pp 11–23

Bendtsen JD, Nielsen H, von Heijne G, Brunak S (2004) Improved prediction of signal peptides: SignalP 3.0. J Mol Biol 340:783–795

Boyington JC, Motyka SA, Schuck P, Brooks AG, Sun PD (2000) Crystal structure of an NK cell immunoglobulin-like receptor in complex with its class I MHC ligand. Nature 405:537–543

Brown DM, Zeef LAH, Ellis J, Goodacre R, Turner SR (2005) Identification of novel genes in Arabidopsis involved in secondary cell wall formation using expression profiling and reverse genetics. Plant Cell 17:2281–2229

Coult NJ, Tisi D, Hohenester E (2003) Novel fold revealed by the structure of a FSA1 domain pair from the insect cell adhesion molecule fasciclin I. Structure (Camb) 11:197–203

Dahiya P, Findlay K, Roberts k, McCann M (2006) A fasciclin-domain containing gene, *ZeFLA11*, is expressed exclusively in xylem elements that have reticulate wall thickenings in the stem vascular system of *Zinnia elegans* cv Envy. Planta 223:1281–1291

Danyluk J, Perron A, Houde M, Limin A, Fowler B, Benhamou N, Sarhan F (1998) Accumulation of an acidic dehydrin in the vinicity of the plasma membrane during cold acclimation of wheat. Plant Cell 10:623–638

Dayhoff MO, Schwartz RM, Orcutt BC (1978) A model of evolutionary change in proteins. In: Dayhoff MO (eds) Atlas of Protein Sequence Structur, vol 5, Suppl. 3. National Biomedical Research Foundation, Washington, DC, pp 345–352

Eisenhaber B, Wildpaner MJ SC, Borner GHH, Dupree P, Eisenhaber F (2003) Glycosylphosphatidylinositol lipid anchoring of plant proteins. Sensitive prediction from sequence- and genome-wide studies for Arabidopsis and rice. Plant Physiol 133:1691–1701

Elkins T, Hortsch M, Bieber AJ, Goodman CS (1990) Drosophila fasciclin 1 is a novel homophilic adhesion molecule that along with fasciclin III can mediates cell sorting. J Cell Biol 110:1825–1832

Felsenstein J (1993) Phylogeny Inference Package (PHYLIP). Version 3.5. University of Washington, Seattle

Fincher GB, Sawyer WH, Stone BA (1974) Chemical and Physical Properties of an Arabinogalactan-Peptide from Wheat Endosperm. Biochem J 139:535–545

Gaspar YM, Nam J, Schultz CJ, Lee LY, Gilson PR, Gelvin SB, Bacic A (2004) Characterization of the Arabidopsis lysine-rich arabinogalactan-protein AtAGP17 mutant (rat1) that results in a decreased efficiency of agrobacterium transformation. Plant Physiol. 135:2162–2171

Gleeson PA, McNamara M, Wettenhall REH, Stone BA, Fincher GB (1989) Characterization of the hydroxyproline-rich protein core of an arabinogalactan-protein secreted from suspension-cultured *Lolium multiflorum* (Italian ryegrass) endosperm cells. Biochem J 264:857–862

Godzik A (2003) Fold recognition methods. Methods Biochem Anal 44:525–546

Guindon S, Lethiec F, Duroux P, Gascuel O (2005) PHYML Online–a web server for fast maximum likelihood-based phylogenetic inference. Nucleic Acids Res 33(suppl 2):W557–W559

Horton P, Park KJ, Obayashi T, Nakai K (2006) Protein subcellular localization prediction with WoLF PSORT. In: Proceedings of the 4th annual Asia Pacific bioinformatics conference APBC06, Taipei, Taiwan, pp 39–48

Huang X, Madan A (1999) CAP3: a DNA sequence assembly program. Genome Res 9:868–877

Johnson KJ, Jones BJ, Bacic A, Schultz CJ (2003) The fasciclin-like arabinogalactan proteins of arabidopsis. A multigene family of putative cell adhesion molecules. Plant Physiol 133:1911–1925

Jones DT, Taylor WR, Thornton JM (1992) The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci (CABIOS) 8:275–282

Kawamoto T, Noshiro M, Shen M, Nakamasu K, Hashimoto K, Kawashima-Ohya Y, Gotoh O, Kato Y (1998) Structural and phylogenetic analyses of RGD-CAP/beta ig-h3, a fasciclin-like adhesion protein expressed in chick chondrocytes. Biochim Biophys Acta 1395:288–292

Kieliszewski MJ, Leykam JF, Lamport DTA (1990) Structure of the threonine-rich extensin from *Zea mays*. Plant Physiol 92:316–326

Kim JE, Jeong HW, Nam JO, Lee BH, Park RW, Kim KS, Kim IS (2000) Identification of motifs for cell adhesion within the repeated domains of the transforming growth factor-beta-induced gene, beta ig-h3. J Biol Chem 275:30907–30915

Kim JE, Kim SJ, Lee BH, Choi JY, Park RW, Park JY, Kim IS (2002) Identification of motifs in fasciclin domains of the transforming growth factor-b-induced matrix protein βig-h3 that interact with the αvβ5 integrin. J Biol Chem 277:46159–46165

Kronegg J, Buloz D (1999) Detection/prediction of GPI cleavage site (GPI-anchor) in a protein (DGPI) retrieved from http://129.194.185.165/dgpi/

Lafarguette F, Leplé J-C, Déjardin A, Laurans F, Costa G, Lesage-Descauses MC, Pilate G (2004) Poplar genes encoding fasciclin-like arabinogalactan proteins are highly expressed in tension wood. New Phytol 164:107–121

Matsuoka K, Watanabe N, Nakamura K (1995) O-glycosylation of a precursor to sweet potato vacuolar protein, sporamin, expressed in tobacco cells. Plant J 8:877–889

Marchler-Bauer A, Bryant SH (2004) CD-Search: protein domain annotations on the fly. Nucleic Acids Res 32:W327–W331

Nothnagel EA (1997) Proteoglycans and related components in plant cells. Int Rev Cytol 174:195–291

Persson S, Wei H, Milne J, Page GP, Somerville CR (2005) Identification of genes required for cellulose synthesis by regression analysis of public microarray data sets. Proc Natl Acad Sci USA 102:8633–8638

Sambrook J, Fritsch EF, Maniatis T (1989) Molecular cloning: a laboratory manual, 2nd edn. Cold Spring Harbor Laboratory Press, Cold Spring Harbor

Schultz CJ, Rumsewicz MP, Johnson KL, Jones BJ, Gaspar YM, Bacic A (2002) Using genomic resources to guide research directions. The arabinogalactan protein gene family as a test case. Plant Physiol 129:1448–1463

Shi H, Kim YS, Guo Y, Stevenson B, Zhu JK (2003) The Arabidopsis SOS5 locus encodes a putative cell surface adhesion protein and is required for normal cell expansion. Plant Cell 15:19–32

Shimizu M, Igasaki T, Yamada M, Yuasa K, Hasegawa J, Kato T, Tsukagoshi H, Nakamura K, Fukuda H, Matsuoka K (2005) Experimental determination of proline hydroxylation and hydroxyproline arabinogalactosylation motifs in secretory protein. Plant J 42:877–889

Showalter AM (2001) Arabinogalactan-proteins: structure, expression and function. Cell Mol Life Sci 58:1399–1417

Soroka V, Kiryushko D, Novitskaya V, Rønn LCB, Poulsen FM, Holm A, Bock E, Berezin V (2002) Induction of neuronal differentiation by a peptide corresponding to the homophilic binding site of the second Ig module of the neural cell adhesion molecule. J Biol Chem 277:24676–24683

Tan L, Leykam JF, Kieliszewski MJ (2003) Glycosylation motifs that direct arabinogalactan addition to arabinogalactan-proteins. Plant Physiol 132:1362–1369

Wang JH, Smolyar A, Tan K, Liu JH, Kim M, Sun Z-J, Wagner G, Reinherz EL (1999) Structure of a heterophilic adhesion complex between the human CD2 and CD58 (LFA-3) counterreceptors. Cell 97:791–803