The family of DOF transcription factors: from green unicellular algae to vascular plants

Miguel Ángel Moreno-Risueno · Manuel Martínez · Jesús Vicente-Carbajosa · Pilar Carbonero

Abstract This article deals with the origin and evolution of the DOF transcription factor family through a phylogenetic analysis of those DOF sequences identified from a variety of representative organisms from different taxonomic groups: the green unicellular alga *Chlamydomonas reinhardtii*, the moss *Physcomitrella patens*, the fern *Selaginella moellendorffii*, the gymnosperm *Pinus taeda*, the dicotyledoneous *Arabidopsis thaliana* and the monocotyledoneous angiosperms *Oryza sativa* and *Hordeum vulgare*. In barley, we have identified 26 different DOF genes by sequence analyses of clones isolated from the screening of genomic libraries and of ESTs, whereas a single DOF gene was identified by bioinformatics searches in the *Chlamydomonas* genome. The phylogenetic analysis groups all these genes into six major clusters of orthologs originated from a primary basal grade. Our results suggest that duplications of an ancestral DOF, probably formed in the photosynthetic eukaryotic ancestor, followed by subsequent neo-, sub-functionalization and pseudogenization processes would have triggered the expansion of the DOF family. Loss, acquisition and shuffling of conserved motifs among the new DOFs likely underlie the mechanism of formation of the distinct subfamilies.

Keywords DOF · Transcription factors · Molecular phylogenetics · Unicellular algae · Vascular plants

Introduction

Gene expression is regulated in living cells through different mechanisms, although the most important one is transcriptional control. DNA-binding transcription factors (TFs) regulate the rate of transcription by interacting, through their DNA-binding domains, with *cis*-regulatory elements (CREs) in the promoters of their target genes. DNA-binding domains are, in general, highly conserved and have been used to classify the TFs into families. Some of these families are common to most eukaryotic organisms but others are specific to a given taxonomic group. A family of TFs putatively specific to plants is the DOF (DNA-binding with One Finger) family (Yanagisawa 2002; Lijavetzky et al. 2003). DOF proteins share a DNA-binding domain of 52 amino acid residues that is structured as a Cys2/Cys2 Zn^{2+} finger (Umemura et al. 2004) that recognizes CREs containing the common core 5′-AAAG-3′ (Yanagisawa and Schmidt 1999). Phylogenetic relationships between rice and Arabidopsis DOF proteins have been previously established (Lijavetzky et al. 2003), and four major clusters of orthologous genes (MCOGs) identified. Phylogenetic analyses have been also described for other plant TF families such as the WRKY, R2R3–MYB, bZIP, MADS, GATA, etc. (Eulgem et al. 2000; Martin and Paz-Ares 1997; Jakoby et al. 2002; Parenicova et al. 2003; Reyes et al. 2004), and some of these analysis have focused on the origin and expansion of a particular TF family. Zhang and Wang (2005) proposed that WRKY genes were originated in early eukaryotes and greatly expanded in plants. Becker and Theissen (2003) reported that MADS diversity is rather ancient, and Shigyo et al. (2006) showed that the AP2 subfamily diverged into the AP2 and ANT groups before the origin of the land-plant lineage. In addition, Maizel et al. (2005) demonstrated that the DNA binding domain of the LEAFY MADS TF, although largely conserved from mosses to angiosperms, has diverged in activity.

DOF TFs have been suggested to participate in the regulation of vital processes exclusive to plants such as photosynthetic carbon assimilation, light-regulated gene expression, accumulation of seed-storage proteins, germination, dormancy, response to phytohormones, flowering time, and guard cell-specific gene expression (Lijavetzky et al. 2003 and references cited therein) that represent important adaptations acquired throughout plant evolution. Therefore, a phylogenetic analysis of this gene family in the representative species of different taxonomic groups appeared through this process seems to be pertinent. In this study, we have performed such an analysis. The search of public databases allowed us the identification of a single DOF gene in the green unicellular alga *Chlamydomonas reinhardtii*, 9 DOF genes in the genome of the moss *Physcomitrella patens*, 8 in the fern *Selaginella moellendorffii*, and 8 in the gymnosperm *Pinus taeda*, besides the 36 genes previously identified in the genome of the dicotyledoneous angiosperm *A. thaliana*, and the 30 genes in the monocot *Oryza sativa* (Lijavetzky et al. 2003). In barley, we have characterized 24 new different genes from BAC and genomic libraries and from ESTs, besides the *Pbf* and *Sad* genes previously identified by us (Mena et al. 1998; Isabel-LaMoneda et al. 2003). The phylogenetic analysis of all these DOFs, using Bayesian inference, shows that they can be grouped into six major clusters of orthologous and paralogous genes that probably originated by gene duplication events from a paraphyletic basal grade (subfamily A). Our results suggest that gain and loss of conserved motifs among the seven subfamilies identified, as well as neofunctionalization, pseudogenization and subfunctionalization processes seem to have occurred throughout the evolutionary process.

Materials and methods

Screening of barley genomic, BAC and EST libraries

A 150 bp probe corresponding to the DOF domain of the barley *Pbf* gene (Mena et al. 1998) was used to screen a *Hordeum vulgare* cv Morex BAC library and a cv Igri genomic library in Lambda FIXII (Stratagene; Isabel-LaMoneda et al. 2003). Eight filters from the BAC Library HV_MBa (Clemson University Genomics Institute; Yu et al. 2000) representing 3.12 times the barley genome were hybridized under standard conditions (Sambrook and Russell 2001). DNA of selected clones, which was isolated through R.E.A.L. Prep 96 Plasmid Kit (QIAGEN), was first sequenced with a specific primer for the DOF domain (5′ ACACCAAGTTCTGCTAC 3′) and subsequently with specific primers for each gene. The genomic library representing 5×10^5 plaque forming units was plated after infection of the *E. coli* strain XL1-blue MRF, the plates were transferred onto Hybond-N membranes (Amersham) and the filters hybridized and washed under standard conditions. Several positive plaques were identified and their nucleotide sequences determined with vector-specific primers. These nucleotide sequences were used to search for ESTs of DOF genes in publicly available libraries, using the BLASTN program (Altschul et al. 1997) at the "National Center for Biotechnology Information" (NCBI) website (http://www.ncbi.nlm.nih.gov/). Selected ESTs were obtained and sequenced in both strands using vector-specific primers. All the DNA sequences were obtained in an automated DNA sequencer (ABI PRISM TM 3100; Perkin Elmer-Applied Biosystems). Sequences of genomic clones, BACs and ESTs were assembled at "The Cap EST Assembler" from IFOM (The FIRC Institute of Molecular Oncology) website (http://www.bio.ifom-firc.it/ASSEMBLY/assemble.html), and non-redundant genes identified through BLASTCLUST within the BLAST package (Altschul et al. 1997) downloaded at NCBI.

Bioinformatic search for new DOF genes

The nucleotide DOF domain sequences of barley *Pbf* and *HvDof1*, rice *OsDof16* (Os03g60630) and Arabidopsis *AtDof3* (At5g60850) genes were used to search for potential DOF genes in the genome of *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Danio rerio*, *Mus musculus*, *Pinus taeda*, *Physcomitrella patens*, *Selaginella moellendorffii*, *Chlamydomonas reinhardtii*, *Cyanidioschyzon merolae*, *Galdieria sulphuraria* and *Emiliania huxleyi* through BLASTN, TBLASTX and discontinuous MEGABLAST (Altschul et al. 1997) at the "Plant Genome DataBase" (http://www.plantgdb.org/), "Chlamy Center" (http://www.chlamy.org/), NCBI and "*Cyanidioschyzon merolae* Genome project" (http://www.merolae.biol.s.u-tokyo.ac.jp/) websites.

Phylogenetic tree

The amino acid sequences of the DOF genes from *C. reinhardtii*, *P. patens*, *S. moellendorffii*, *P. taeda*, *H. vulgare*, *O. sativa*, and *A. thaliana* were deduced through the "Translate tool" at ExPASy Proteomics Server (http://www.expasy.ch/). The identification of a homologous region in all these protein sequences that spanned the classical DOF-binding domain plus three amino acid residues at its end was performed through a multiple alignment using CLUSTAL W (Thompson et al. 1997). The alignment of these homologous regions prior to the phylogenetic analysis was also carried out by means of CLUSTAL W. A tree was inferred by Bayesian phylogenetic inference using MrBayes v3.1 (Huelsenbeck and Ronquist 2001). Four Markov chains (default temperature) starting with a random tree were run simultaneously for 15,000,000 generations with sampling from the trees made every 1,500th generation under the Dayhoff model (Dayhoff et al. 1978). Altogether 2,500 trees (out of 10,000) were discarded as the burn-in based on the stationary phase. Each analysis was done twice, and the final tree was computed from the combination of accepted trees from each analysis. Convergence between the two runs was tested after examination of the potential scale reduction factor (PSRF), which was 1.000, and the standard deviation of split frequencies, which was below 0.01. Robustness of the inferred tree was evaluated using Bayesian posterior probabilities.

Identification of conserved motifs

The deduced protein sequences of the 116 DOF genes from *C. reinhardtii*, *P. patens*, *S. moellendorffii*, *P. taeda*, *H. vulgare*, *O. sativa* and *A. thaliana* were analyzed by means of the MEME program (http://www.meme.sdsc.edu/meme/ meme.html) as described by Bailey and Elkan (1994). Default parameters were used with the following exceptions: the occurrence of a single motif was set to any number of repetitions, the maximum number of motifs to find was set to 60 and the minimum width of each motif was set to eight amino acid residues.

Shown consensus sequences follow the criteria of Joshi et al. (1997): a single capital letter is given if the relative frequency of a single residue at a certain position is greater than 50% and greater than twice that of the second most frequent residue. When no single residue satisfies these criteria, a pair of residues is assigned as capital letters in brackets if the sum of their relative frequencies exceeds 75%. If neither of these criteria is fulfilled, a lower-case letter is given if the relative frequency of a residue is greater than 40%. Otherwise, x is given. Conserved motifs identified by MEME were scanned using WoLF PSORT (Nakai and Horton 1999) at PSORT server (http://www.psort.org/) to identify subcellular localization signals.

Comparison of sequences to determine the exon–intron structure of the genes

EST sequences and JGI gene models vs 2.0 obtained at the "Chlamy center" (http://www.chlamy.org/) besides *H. vulgare* and *P. patens* genomic and EST/cDNA sequences were compared through EMBOSS Pairwise Alignment Algorithm—method needle (global)—at "European Bioinformatics Institute" website (http://www.ebi.ac.uk/emboss/align/). Exon–intron structure of some genes from *H. vulgare* was inferred from comparison with their putative orthologous genes from *O. sativa*, and those of *SmDof3* and *SmDof4* by mutual comparison through the same algorithm. The exon–intron structure of the *A. thaliana* and *O. sativa* genes is that of the TAIR (The Arabidopsis Information Resource) website (http://www.arabidopsis.org/) and "The TIGR Rice Genomic Annotation Project" (http://www.tigr.org/tdb/e2k1/osa1/).

Expression pattern and duplicated DOF genes in *A. thaliana*

Processed absolute signal intensity values (linear data) from ATH1_22K array were obtained by means of the META-ANALIZER tool at the GENEVESTIGATOR program (Zimmermann et al. 2004; http://www.genevestigator.ethz.ch/). Data corresponding to developmental stages based on the key ontology from Boyes et al. (2001) were normalized for gray scale such that the signal corresponding to intensity 1500 was assigned the value 100% (black) and absence of signal 0% (white). For data corresponding to response to light, a "+/−" mark was assigned if the response of the gene under treatment was, respectively, higher/lower than the signal intensity of the control. DOF genes in *A. thaliana* duplicated genomic regions were identified at "MATDB: Redundancy Viewer" and "MATDB: Segmental Duplications" from MIPS (Munich Information Center for Protein Sequences) (http://www.mips.gsf.de/projects/plants/). Data for duplicated DOF genes were normalized for gray scale such that the maximum value intensity reached for each couple of duplicates was assigned 100% (black).

Results

Search for new DOF genes: only positive matches were found in plants and plant-related organisms

We started to search for members of the DOF family in the yeast *Saccharomyces cerevisiae*, the nematode *Caenorhabditis elegans*, the insect *Drosophila melanogaster*, the fish *Danio rerio*, and the mammalian *Mus musculus*, whose genomes have been completely sequenced. The negative result of such a search indicated that DOF genes were unlikely to exist in these eukaryotes.

Then, we explored several photosynthetic eukaryotes from different clades originated during evolution: the green unicellular alga *Chlamydomonas reinhardtii*, the moss *Physcomitrella patens*, the fern ally *Selaginella moellendorffii*, the gymnosperm *Pinus taeda*, the dicotyledoneous angiosperm *A. thaliana* and the monocotyledoneous angiosperms *O. sativa* and *H. vulgare*. After a thorough examination, only a single DOF gene (*CrDof1*) was found in *C. reinhardtii* (supplementary material S1), whereas searches for DOF genes in other algae not belonging to the viridiplantae (*Cyanidioschyzon merolae*, *Galdieria sulphuraria* and *Emiliania huxleyi*) did not result in any positive match. Searches in the genomes of *P. patens* and *S. moellendorffii* allowed us to identify nine non-redundant genes in the moss (*PpDof1 to PpDof9*; S1) and eight non-redundant genes in the fern (*SmDof1 to SmDof8*; S1). In the gymnosperm *P. taeda* eight non-redundant genes were found (*PtDof1 to 8*, S1).

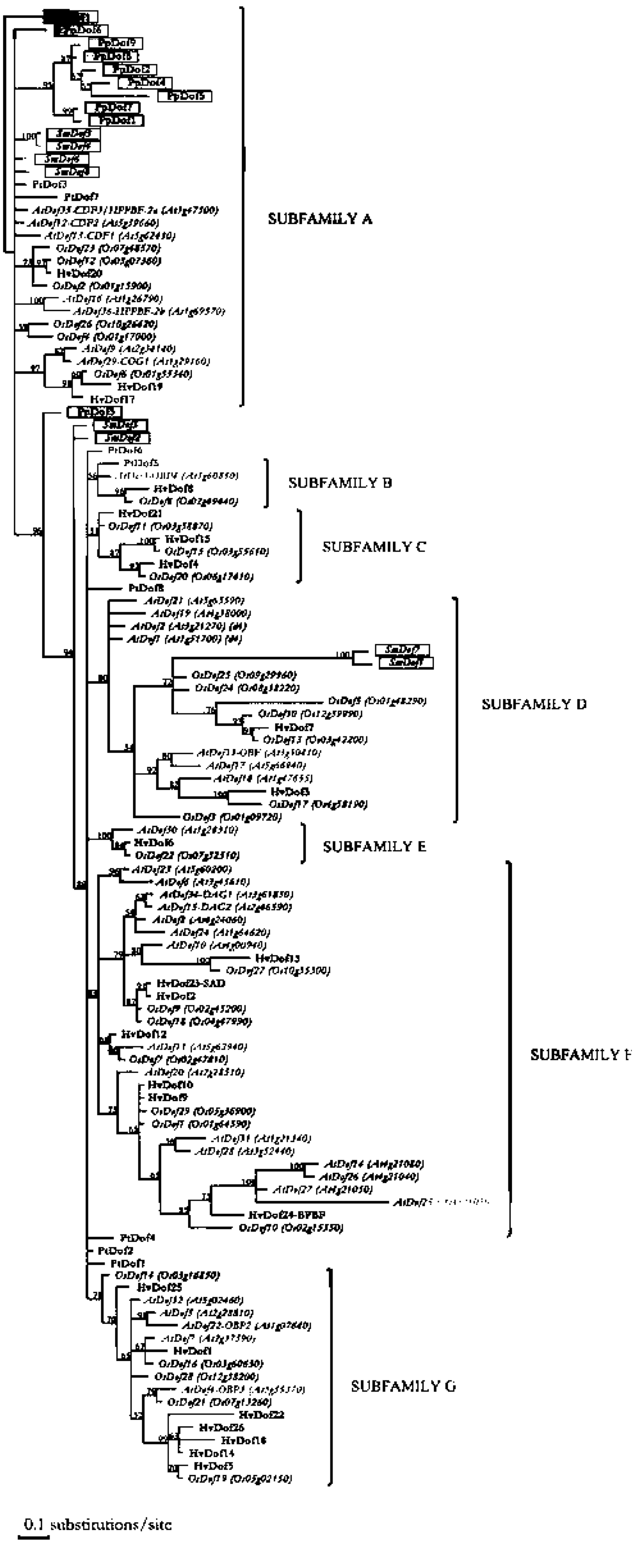
In our search for DOF proteins in the barley genome that is not still completely sequenced, we screened BAC and genomic λ-based libraries and looked for ESTs in publicly available databases. From 20 λ positive clones isolated from the genomic library, only eight non-redundant genes were found: barley *Pbf* and *Sad* (BPBF and SAD proteins, reassigned *HvDof24* and *HvDof23*, respectively), *HvDof19* to *22* and *HvDof25* to *26* (S1). In the BAC library, the 76 positive clones identified corresponded to 23 non-redundant genes, from which 16 turned out to be new DOF genes (*HvDof1* to *16*). The positive matches identified in our search of ESTs databases were ordered, sequenced and compared to render two new DOF genes (*HvDof17* and *HvDof18*). In addition, the already described 36 DOF proteins from *A. thaliana* (*AtDof1-36*), and the 30 from *O. sativa* (*OsDof1-30*) were considered in our analysis (Lijavetzky et al. 2003).

Phylogenetic relationships

The phylogenetic relationships among the DOF genes previously identified were inferred from the derived amino acid sequences of the genes from *C. reinhardtii*, *P. patens*, *S. moellendorffii*, *P. taeda*, *H. vulgare*, *O. sativa*, and *A. thaliana*. These 118 proteins were aligned and a common homologous region identified. This region spanned 55 amino acid residues that included the 52 amino acids of the previously described DOF DNA-binding domain. Since the DOF domains of *HvDof11* and *HvDof16* were truncated, these genes were not included in the subsequent phylogenetic analysis. Bayesian Inference was applied to the alignment of the identified homologous region of the other 116 deduced proteins (S2), and a phylogenetic tree reconstructed (Fig. 1). The only DOF from *C. reinhardtii* was used to root the tree, assuming that *C. reinhardtii* and the other species considered here evolved from a common ancestor. The tree topology and the corresponding phylogenetic relationships indicated that the majority of the DOF proteins analyzed, independently of the species, could be grouped into seven subfamilies (A–G) or major clusters of orthologous genes (MCOG). The A subfamily constituted a paraphyletic basal grade containing the *C. reinhardtii* only gene, as well as DOF genes from the other species. The other six subfamilies (B–G) were located in a branch posterior to the basal grade and supported by Bayesian probabilities higher than 70% (D–G) or 50% (B and C). Eight out of the nine DOF genes from the bryophyte *P. patens* belonged to subfamily A, while *PpDof3* was in an intermediate branch between this subfamily and the branch in which the other six subfamilies (B–G) were located. Four *S. moellendorffii* genes were located in the initial grade, two in an intermediate branch between subfamily A and the rest of the subfamilies, and the other two in subfamily D. Two DOFs from the gymnosperm *P. taeda* belonged to the basal grade (*PtDof3* and *7*), one to subfamily B (*PtDof5*), and the other five were located in the same branch that anchored subfamilies B to G (*PtDof1*, *2*, *4*, *6* and *8*). The DOF genes from the two monocot species, *H. vulgare* and *O. sativa* appeared in all the subfamilies, and subfamily C was only integrated by genes from these species. Therefore, it seems that consecutive gene duplications would have triggered a progressive expansion from subfamily A.

Acquisition and loss of amino acid motifs in DOF proteins in the course of evolution

The analysis of the complete amino acid sequences of the DOF proteins, using the MEME software, revealed the existence of homologous motifs, conserved among their sequences and different from the DOF binding domain characteristic of this family (Table 1). These



◀**Fig. 1** Phylogenetic tree of the DOF family in seven representative species of some of the main viridiplantae lineages. DOFs from the green algae *C. reinhardtii* are *boxed* in *dark gray*, from the moss *P. patens* are *boxed* in *pale gray*, from the fern ally *S. moellendorffii* are in *italics* and *boxed* in *pale gray*, from *P. tadea* are in *gray*, from *H. vulgare* are in *black*, from *O. sativa* in *dark gray* and *italics*, and from *A. thaliana* in *pale gray* and *italics*. The rooted tree was inferred in MrBayes v3.1 by the Dayhoff model over the alignment of a common homologous region spanning 55 amino acids that included the 52 amino acids of the DOF domain of the 116 DOFs from the previously mentioned organisms. The resulting major clusters of orthologous genes (MCOGs) or subfamilies are shown on the *right*. The *scale bar* corresponds to 0.1 estimated amino acid substitutions per site. Bayesian posterior probabilities are indicated

motifs seemed to be specific to the different subfamilies rather than representative for each species. To produce a schematic view of the common motifs found in the majority of the proteins, we only considered those motifs present in at least 50% of the sequences of a given species that were also shared by proteins from some of the other species analyzed. These motifs were represented in their relative location within the protein (Fig. 2a).

This schematic evolutionary tree showed that the specific motifs present in subfamily A were not present in the other subfamilies, with the exception of motif 5 that was also found in one of the DOFs from *P. patens* (*PpDof3*) located in a branch between subfamily A and subfamilies B–G. Moreover, this protein from *P. patens* also had motif 10, which was shared by proteins belonging to subfamilies B to G. The proteins from *S. moellendorffii* connecting subfamilies A to B-G shared motif 9 with proteins from B and F subfamilies and with some from *P. taeda* located in the same branch as subfamilies B–G. Motif 19, which was characteristic of the F subfamily, and motif 8, specific to the G subfamily, were also present in some of these mentioned proteins from *P. taeda* located in the branch anchoring subfamilies B–G. In addition, some motifs were only found in a given subfamily, as was the case of motifs 8, 7 and 15 from subfamily G or motifs 12, 17, 14, 29, 13 and 26 from subfamily F. Finally, transpositions seemed to have occurred: (1) motifs 27 and 22 appeared exchanged in their relative position along the C-terminal part of the proteins from *C. reinhardtii* and *P. patens*, and (2) motif 6 had changed from an initial position in the C-terminal part in *P. patens* to an N-terminal region in angiosperms.

The sequence similarity of motifs 5, 9 and 3 (Fig. 2b) and their close position to the DOF domain could be indicative of a common origin and an associated functionality. Representative consensus amino acid sequences of these motifs in each species were obtained. Motif 5 was detected in the green alga *C. reinhardtii* as well as in the rest of the species considered in this study. This motif 5 might have originated in common with a related motif 5′ present in *P. patens* that shared a high similarity to motif 9 of *S. moellendorffii* and *P. taeda*. Motif 9 was also present in the angiosperms *A. thaliana*,

Table 1 Conserved amino acid motifs obtained by means of MEME (Bailey and Elkan 1994) from the analysis of the 116 DOF protein sequences of *C. reinhardtii*, *P. patens*, *S. moellendorffii*, *P. taeda*, *H. vulgare*, *O. sativa* and *A. thaliana*

Motif	*E*-value	Consensus sequences
1	8.3e−484	C-P-R-C-x-S-x-x-T-K-F-C-Y-Y-N-N-Y-[SN]-l-x-Q-P-R-[HY]-F-C-[KR]-x-C-x-R-Y-W-T-x-G-G-x-L-R-N-V-P-[IV]-G-G-G-x-R-K
2	1.4e−151	D-x-K-x-x-g-x-L-W-V-P-K-T-L-R-I-D-D-P-[DN]-E-A-A-K-S-S-I-W-[ST]-T-L-G-I-K-x-d-[DK]-x-[AG]-x-[FD]
3	9.3e−148	P-x-S-M-s-E-R-A-R-[LM]-A-[RK]-[IV]-P-x-P-E-[PQ]-G-L-[KN]
4	6.3e−084	[DG]-D-p-g-I-K-L-F-G-[KR]-[TV]-I-[PT]-f
5[a]	1.3e−072	[KT]-t-x-x-d-s-x-x-k-x-x-L-K-K-P-D-K-I-L-P
6	7.0e−058	[VA]-x-e-[TS]-[LS]-P-x-L-x-A-N-P-A-A-[FL]-S-R-S-x-[NS]-F-x-E-[ST]-[ST]
7	4.9e−038	L-E-Q-W-R-[AL]-[AQ]-Q-M-[EQ]-S-F-P-F-[FL]-H-A-M-D-H-Q
8	2.2e−029	M-V-F-S-S-V-[PQ]-x-[CY]-[LM]-D-[PS]-[PS]-[ND]-W
9[a]	7.5e−024	E-R-R-a-R-P-q-k-[DE]-Q-[AP]
10	3.3e−025	N-K-R-S-x-S-x-S
11	7.2e−025	E-V-V-D-E-S-R-D-E-E-A-I-T-E-D-D-H-S-P-P-A-E-Q-E-E-V-I-E-V-V-D-H-Q-E-E-D-Y-C-D-I-D
12	9.8e−012	[GI]-n-V-K-P-M-E-E-[IM]-x-[MT]
13	1.4e−021	R-[LM]-L-F-[AP]-F-E-D-L-K
14	3.9e−019	G-A-L-S-A-M-E-L-L-R-S-T-G-C-Y-M-P-L-Q-[MV]
15	5.0e−020	L-[IL]-[AT]-Q-[LM]-A-S-V-K-M-[DE]-[DE]-[NS]
16	3.9e−024	[SY]-S-[SP]-[SP]-x-x-x-[ST]-[SP]-x-S-x-[NS]-[NS]-S-x-[TC]-L-G-K-H-[PS]-R-[DE]-[ES]-[DT]-[DE]-x-e
17	4.5e−015	I-D-L-A-[LM]-L-Y-[AS]-K-F-L
18	5.9e−015	P-P-P-A-P-I-F-A-D-Q-A-A-A-L-A-S-L-F-A-P-P-P-P-P-P
19	1.9e−014	H-E-G-A-x-D-L-N-L-A-F-P-H-H-[HG]
20	7.0e−005	M-Q-E-F-Q-[PS]-I-P-G-L
21	3.5e−011	T-V-A-D-M-[TA]-P-F-[MT]-S-L-D-A-G-I-F-E-L-G-D-[VA]-[PS]-P-A-[AD]-Y-W-N-[AG]-G-S-C-W-T-D-V-[PQ]-D-P-[NS]-V-Y-L
22	6.6e−010	Q-Q-L-S-M-L-P-P-[ST]-G-G-M-L-G-F-G-[DN]-Q-x-S-x-G-x-A-[GP]-[AL]-x-[LP]-P-x-[PS]-[HY]-L-[AQ]-[FL]-[AE]
23	2.1e−007	N-[AP]-s-F-Y-P-v-P-a-Y-W-[GS]-C-p-[IV]
24	2.3e−007	V-S-A-D-R-S-P-N-V-A-Q-H-P-C-M-N-G-G-A-M-W-P-F-G-V-A-P-P-P-A-Y-Y-T-S-S-I-A-I-P-F-Y
25	3.8e−007	E-T-[PV]-x-[EK]-F-[GI]-[PS]-[ED]-[VS]-P-L-C-E-S-M-A-S-V-L-N-I-G-E-Q
26	6.4e−005	G-[FL]-W-[NS]-G-M-I-[GS]
27	1.6e−007	L-H-x-P-S-x-S-D-A-L-A-S-V-M-A-K-A-x-L-W-S-N-P-A-Q-Q-L-P-S-L-N-G-x-T-Q-x-Q-S-S-W
28	6.8e−003	L-W-x-C-x-x-x-W-P-N-G-A-W-N-x-P-W-I
29	2.1e−004	[NL]-L-G-F-S-L-D-x-H-G
30	3.6e−004	E-s-x-l-x-E-E-x-d-e-E-x-[DE]-E

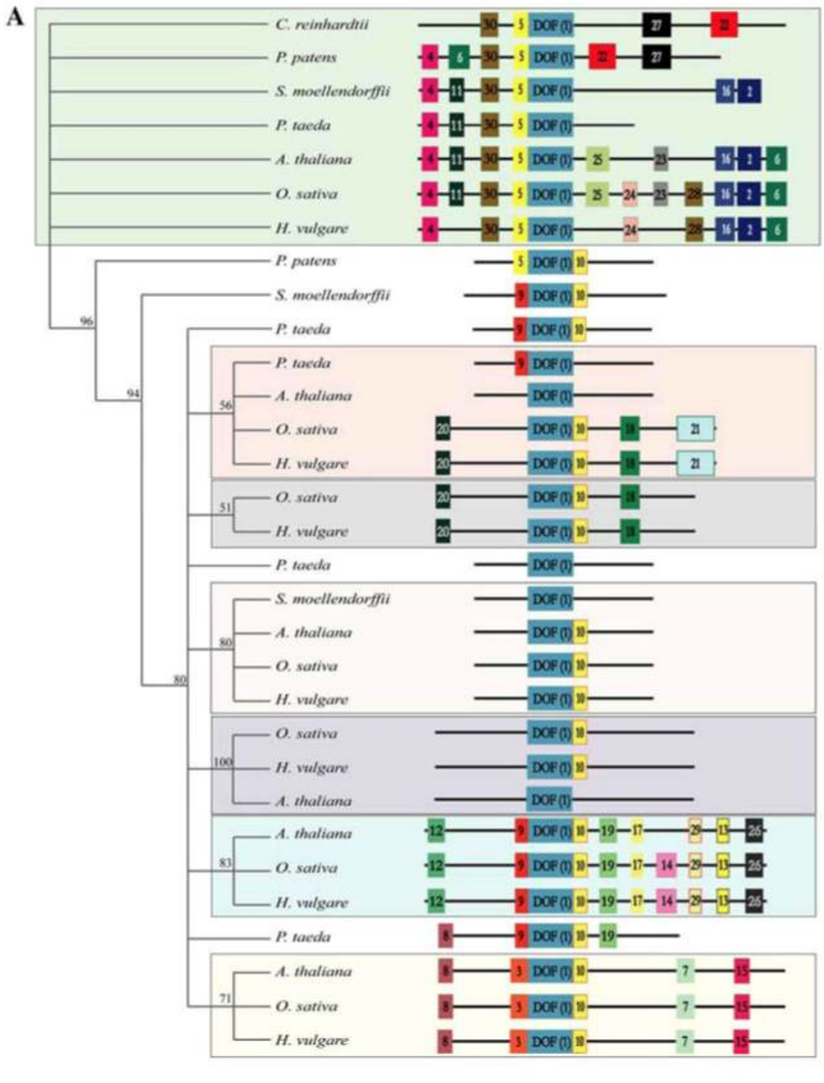
Consensus sequences follow the criterion of Joshi et al. (1997)

[a] These sequences present a putative nuclear localization signal

O. sativa and *H. vulgare* and shared a significant similarity to motif 3. Thus, it is plausible that these motifs associated to the DOF domain and bearing a nuclear localization signal (NLS; motifs 5 and 9, Table 1) could have originated from a common ancestral motif.

Position and homology of introns

The exon-intron structure of *CrDof1* comprised four introns and five exons. This was an unusually complex structure as most of the DOF genes studied here had either one intron or none (Fig. 3). Another feature unique to this gene was that the codon triplet encoding the fourth Cys of the DOF domain, which is indispensable to form the zinc finger, was interrupted by the third intron from the rest of the nucleotides encoding this DNA-binding domain. In addition, the nucleotide region corresponding to motif 5 (Table 1), which contained a putative NLS predicted by PSORT, was interrupted by the second intron. The second and third intron of *CrDof1* from *C. reinhardtii* (Fig. 4) did not have a counterpart in any of the other DOF genes annotated. Regarding the first intron, only nine genes (*OsDof9*, *OsDof10*, *HvDof24-BPBF*, *PpDof7* and *PpDof1* to *PpDof5*) out of the 100 whose genomic structures were established, had an intron in the 5′ non-coding region. Introns found in *P. patens* could be the counterparts of this *CrDof1* intron. However, the introns of the cereal genes seemed to have been acquired later as they are not



B

	MOTIF	ALIGNMENT
C. reinhardtii	5	---------PKLP-R-PDKKEA
P. patens	5	---------AIRE-K-PVKVIP
S. moellendorffii	5	---------QKLK-K-PDKPVA
P. taeda	5	---------KMLK-K-PDKLLP
A. thaliana	5	---------TALK-K-PDKILP
H. vulgare	5	---------EKLK-K-PDKILP
O. sativa	5	---------KKLK-K-PDKILP
P. patens	5´	---------KRLK-P-PDQVIA
S. moellendorfii	9	---------RRLK-PHPDQVL-
P. taeda	9	---------RRSK-PHPTQVQ-
P. taeda	9	---------RRAR-PHPSQVL-
O. sativa	9	---------RRAR-PQKEKAL-
H. vulgare	9	---------RRAR-PQKEKAI-
A. thaliana	9	---------RKAR-PEKDQAL-
P. taeda	9	---------RRLK-PQPEQAL-
O. sativa	3	PVSMAERARLARIPQPEPGLK
H. vulgare	3	PMSMSERARLARVPQPEPGLK
A. thaliana	3	PGSMAERARLANVPLPEAALK

□ Subfamily A	□ Subfamily B	□ Subfamily C	□ Subfamily D
□ Subfamily E	□ Subfamily F	□ Subfamily G	

Fig. 2 **a** Phylogenetic cladogram that includes a schematic representation of the consensus proteins from each species carrying the most representative motifs from those identified by means of MEME software and listed on Table 1. **b** Alignment of the consensus amino acid sequences of motifs 5, 9 and 3 in each species by means of CLUSTAL W. Conserved proline (*P*), lysine (*K*) and arginine (*R*) through evolution are in *blue*, *red*, and *green*, respectively

present in any of the dicot species. In the case of the fourth intron of *CrDof1*, only one gene (*AtDof34-DAG1*) had an intron downstream of the DOF domain, which probably represents a gain of this gene. In general, the presence of introns was restricted to some subfamilies (A, F, and G). Three subfamilies presented genes without introns (B, C, and E), and subfamily D had only two genes

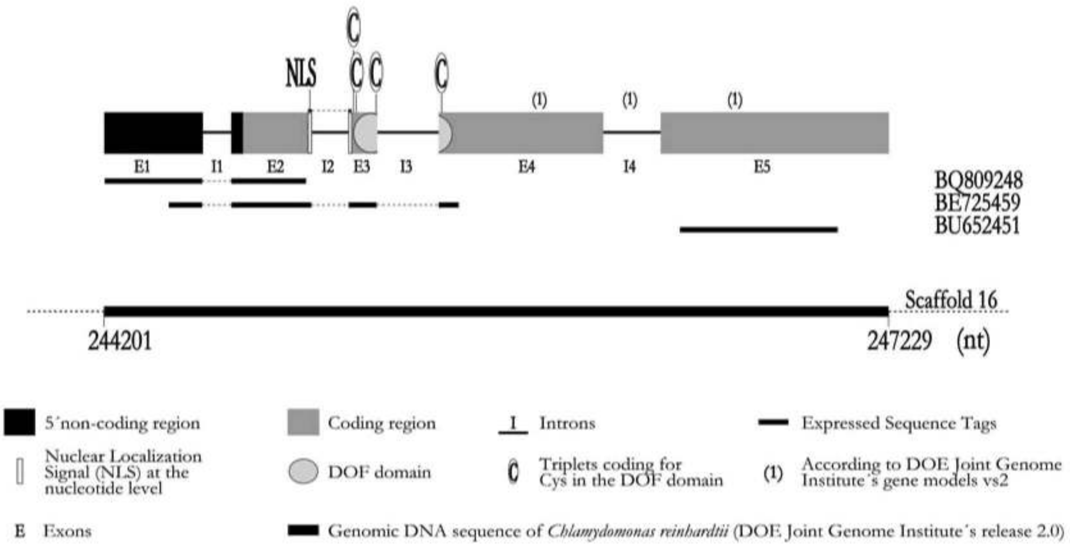

NLS
C
C C
C
(1)
(1)
(1)
E1
I1
E2
I2
E3
I3
E4
I4
E5
BQ809248
BE725459
BU652451
Scaffold 16
244201
247229 (nt)
5´non-coding region
Coding region
Introns
Expressed Sequence Tags
Nuclear Localization Signal (NLS) at the nucleotide level
DOF domain
Triplets coding for Cys in the DOF domain
According to DOE Joint Genome Institute´s gene models vs2
E Exons
Genomic DNA sequence of Chlamydomonas reinhardtii (DOE Joint Genome Institute´s release 2.0)

Fig. 3 Exon–intron structure of the DOF gene from *C. reinhardtii* (*CrDof1*). Over the genomic DNA sequence—scaffold 16 according to the DOE Joint Genome Institute's full DNA sequence release vs2—are shown the ESTs and DOE-JGI's gene model predictions that support this structure

(*HvDof3* and *OsDof17*) with an intron in their genomic structure (Fig. 4). Although intron size was variable, the position of the introns of Arabidopsis, rice and barley putative DOF orthologs was conserved. Genes located in duplicated genomic regions from *A. thaliana* according to MATDB (couples d1–d8), conserved the same exon–intron genomic structure except in the cases d2 and d6, in which one of the genes had more introns than its partner (Fig. 4).

A common expression pattern is not shared by genes belonging to a given DOF subfamily in Arabidopsis

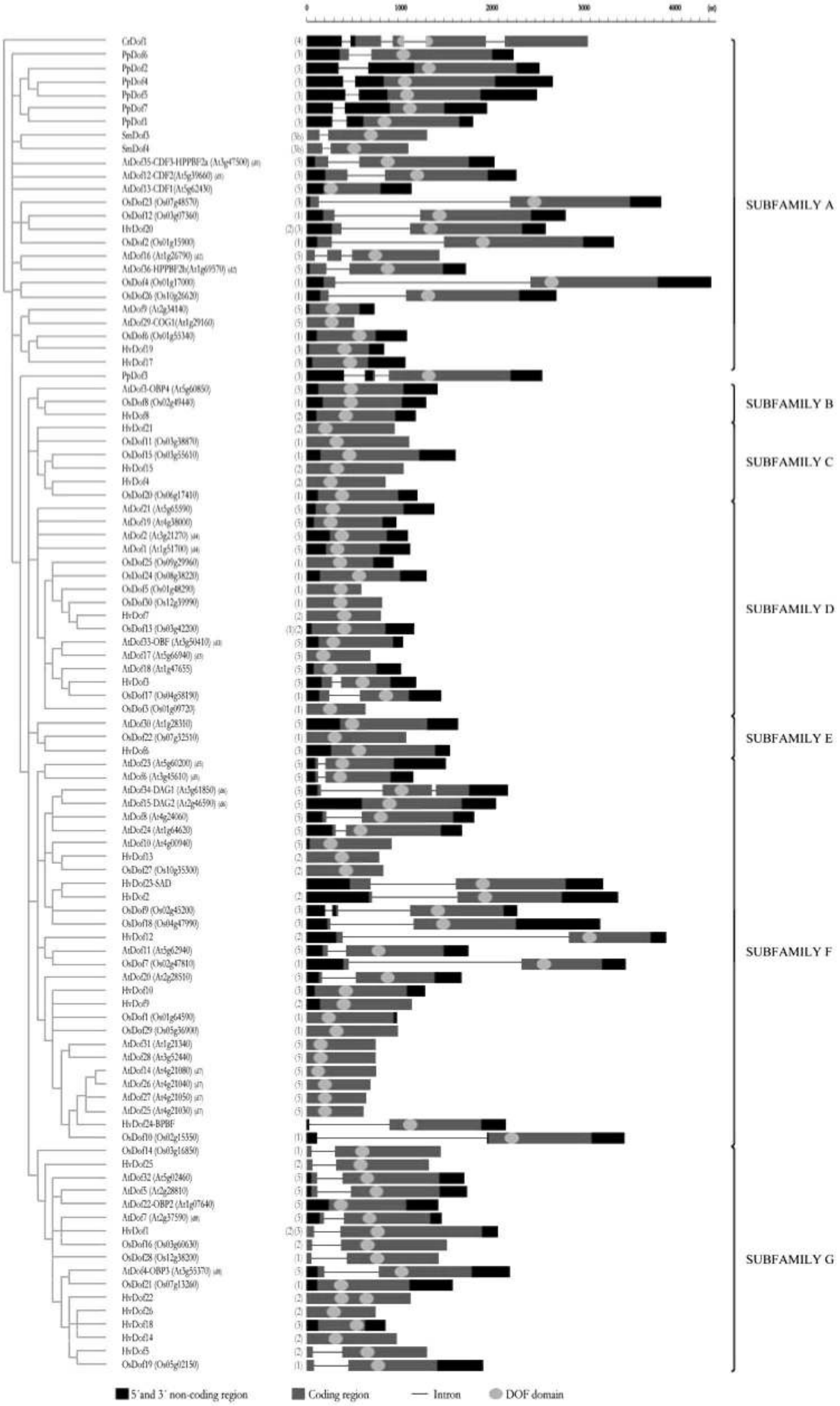
To determine whether the expression patterns of the different DOF genes from *A. thaliana* were characteristic of a given subfamily, data from GENEVESTIGATOR corresponding to hybridizations carried out with the ATH1_22K array in representative developmental stages of the Arabidopsis life cycle and in response to light were compiled. Both, the developmental data, normalized to gray scale, and the induction (+) or repression (-) responses to the different light treatments demonstrated the absence of an obvious expression pattern for each of the subfamilies (Fig. 5). The expression profiles of those Arabidopsis genes located in duplicated genomic regions according to MATDB (couples d1–d8) were partially redundant in the majority of the cases (e.g. d4, d7, d6 and d8). To assess whether this partial redundancy was also detectable in the different tissues, processed data from GENEVESTIGATOR were normalized and couples of duplicated genes d1–d8 were arranged in function of their higher or lower redundant expression (see supplementary material S3). This analysis confirmed that most of the couples also had a redundant expression pattern in the different tissues. *AtDof15* and *AtDof34* (DAG1/DAG2) exhibited the most redundant expression pattern, whereas *AtDof1* and *AtDof2* were the genes with the most complementary expression. In addition, a very low expression was detected for the *AtDof16* gene, whereas its duplicated partner *AtDof36* -HPPBF-2b was strongly expressed in the same tissues.

Discussion

The DOF TFs: an expanding family in the course of evolution

The present phylogenetic study of DOF TFs indicates that this family might have originated from a common ancestor, maintained as a single copy gene in *Chlamydomonas*, and expanded in the different taxonomic groups of vascular plants by recurrent duplication events in the course of evolution. Our phylogenetic tree comprises a paraphyletic grade (subfamily A) containing genes from the 7 representative species analyzed and six subfamilies with genes from some of these species, particularly from angiosperms, but without genes from the moss *Physcomitrella*. In addition, one gene from this moss and two from the fern *Selaginella* (*PpDof3* and *SmDof2* and *5*) are in branches connecting subfamily A to subfamilies B-G, and five from the gymnosperm (*PtDof1*, *2*, *4*, *6* and *8*) are in the branch

Fig. 4 Exon–intron structure of *C. reinhardti*, *P. patens*, *S. moellendorffii*, *A. thaliana*, *O. sativa* and *H. vulgare* DOF genes. The exon–intron structure of *O. sativa* genes is that predicted by TIGR Rice Genome Annotation Project when indicated *(1)*. For *H. vulgare*, *P. patens* and *S. moellendorffii* DOF genes, it has been deduced by comparison with the respective orthologous genes *(2)*, ESTs/cDNAs *(3)* or by mutual comparison *(3b)*. For *C. reinhardtii* it was obtained as mentioned in Fig. 3 *(4)*, and for *A. thaliana* they are those from the TAIR website *(5)*. *(d1–d8)* Genes in duplicated genomic regions according to MATDB redundancy viewer from MIPS



anchoring these subfamilies B to G. Therefore, our analysis supports that the expansion of the DOF family in the course of evolution would have occurred from subfamily A, which constitutes a basal grade, to the rest of the subfamilies. Our results also show that DOF genes are in a variable number that tends to increase as

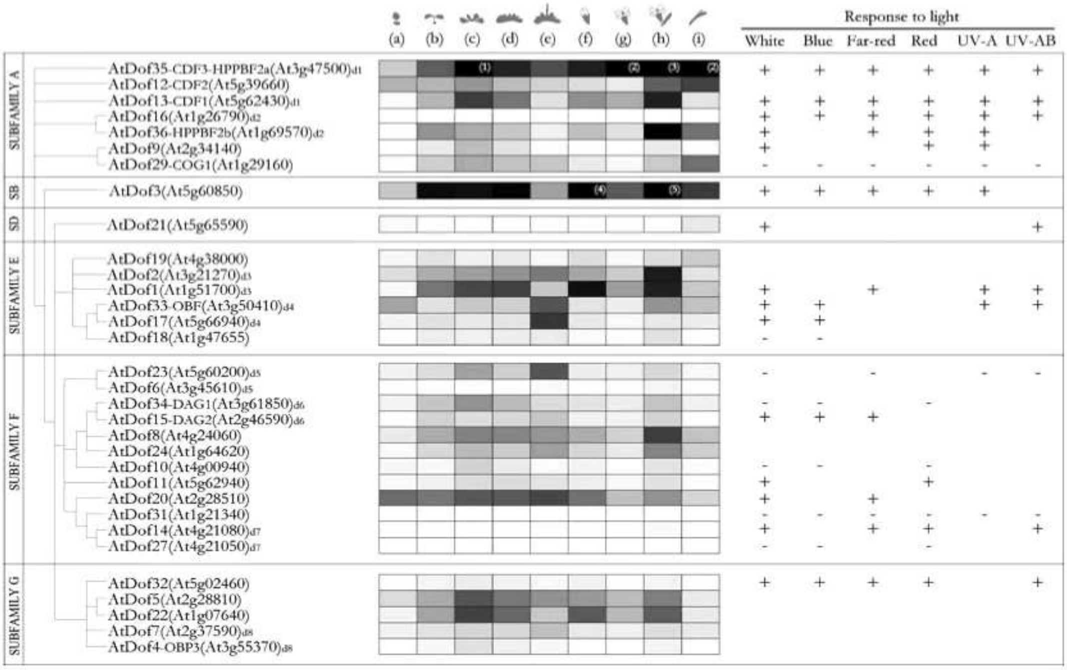


Fig. 5 Expression pattern of the different subfamilies of DOF genes in *A. thaliana*. Data from GENEVESTIGATOR (Zimmermann et al. 2004) were normalized for *gray* scale such that the value 100% was assigned to the signal corresponding to intensity 1500 (*black*) and the 0% to the absence of signal (*white*). (*1*) 199%, (*2*) 105%, (*3*) 176%, (*4*) 111% and (*5*) 122%. A "+"/"−" mark was assigned if the response of the gene under treatment was, respectively, higher/lower than the signal intensity of the corresponding control. Development stages are those in GENEVESTIGATOR that are based on the key ontology from Boyes et al. (2001): (*a*) seed germination, (*b*) cotyledons fully opened, (*c*) 1–5 rosette leaves, (*d*) 6–8 rosette leaves, (*e*) 9–12 rosette leaves, (*f*) 13 rosette leaves—1st flower buds visible, (*g*) 1st flower buds visible—10% of the flowers to be produced have opened, (*h*) 30–50% of the flowers to be produced have opened and (*i*) flowering complete—1st silique shattered. (*d1–d8*) Genes in duplicated genomic regions according to MATDB redundancy viewer from MIPS

the complexity of the organisms do. From the single DOF gene present in the green alga and from 8 or 9 in the gymnosperm and lower plants studied, 26 and 30 appeared in the monocotyledoneous angiosperms and 36 in the dicotyledoneous angiosperm. DOF genes incorporated into these genomes could have acquired new functions, probably regulating transcription of other genes gained in the evolutionary process.

On the origin of DOF genes

The structural features of *CrDof1* could resemble those of the first ancestral DOF gene. This is the only case in a DOF gene in which the nucleotide regions encoding both a putative NLS and the DNA-binding domain were interrupted by introns. Introns are thought to be primordial structures in eukaryote-prokaryote ancestors that would have facilitated the evolution of protein families through exon shuffling (Doolittle 1978; Zhaxybayeva and Gogarten 2003). The fourth Cys of the DOF domain (exon 4) is necessary for the DNA binding of this Cys2/Cys2 zinc finger (Umemura et al. 2004), and exon 3 has certain similarities to SII TFs (Kettenberger et al. 2003) according to searches performed in the PFAM website. Hence, exon shuffling in the photosynthetic eukaryotic ancestor after chloroplast acquisition could have put independent subdomains with different functions together generating a new type of protein with both DNA binding capacity and a NLS. In fact, subdomain recombination has been demonstrated to be able to generate proteins with new properties (Hopfner et al. 1998).

Simplification of complex early eukaryotic gene structures has dominated subsequent evolution. Intron losses outnumber intron gains over a large range of eukaryotic lineages (Roy and Gilbert 2005). In accordance to this, a simplification of the structure of the original DOF, here represented by *CrDof1*, would be expected. However, the presence of introns in some angiosperm DOF genes must have been more recently gained as they were mainly located within subfamilies F and G and positioned in conserved regions of several putative orthologous genes from the two cereal species.

Gene diversification and functionality

As reported for other protein families (Pinyopich et al. 2003), the establishment of new gene functions by duplication events (Long and Langley 1993) contributes to the evolutionary diversification of genomes and is a source of evolutionary novelty (Gilbert et al. 1997). The Ohno's classic model (Ohno 1970) concerning the fate of duplicated genes and the duplication–degeneration–complementation (DDC) model, predict for each one of the duplicates the gain of a new function (neofunctionalization), its loss (pseudogenization) or the development of overlapping redundant functions and expression patterns (subfunctionalization; Force et al. 1999; Lynch and Force 2000). To trace diversification and functionality of DOF genes, *Arabidopsis* represents a model system for which both genome structure and gene expression patterns have been extensively studied. The Arabidopsis plant has suffered different large-scale duplication events (Vision et al. 2000). These events affecting DOF genes would have triggered a process of divergence between the duplicated genes. As shown in Fig. 5 and S3, pairs of duplicated genes d1 and d3 to d8 had partially redundant expression patterns. However, DAG1 and DAG2 (d6), which exhibited the most redundant expression, develop opposite regulatory actions as they repress/promote, respectively, germination in response to light and cold (Gualberti et al. 2002). This effect would be more related to a case of neofunctionalization despite their redundant expression. In addition, a pseudogenization process might be occurring in another pair of duplicated genes, *AtDof16* and *AtDof36*-HPPBF-2b, which is a gene specifically regulated by PhyA in response to far-red light (Tepperman et al. 2001). The former seems to have a noticeably weaker expression than the latter. However, the fact that *AtDof16* responds to light could mean that the pseudogenization has not been completed.

Independently of these neofunctionalization and pseudogenization processes of the newly gained DOF genes in the emerging species, a subfunctionalization of the ancestral DOF role would be expected. Therefore, more than one DOF could be necessary to fully achieve the original function in complex multi-cellular organisms. *CrDof1* is so far the only DOF gene found in an organism different from land plants, but close to the origin of the evolutionary divergence of this lineage. Although the function of *CrDof1* is unknown, it is worth considering that induction of zygote germination in *C. reinhardtii*, a process that emulates seed dormancy, does not occur in the absence of light (Gloeckner and Beck 1995), and that Arabidopsis knockout plants for *AtDof34*-DAG1 germinate in the absence of light (Papi et al. 2000). Thus, a possible role in zygote germination could be anticipated for the Chlamydomonas DOF, although other functions could not be excluded.

On the basis of the phylogenetic tree and on the conserved amino acid motifs, our results allow some conclusions to be drawn concerning the evolution of DOF genes. The presence of Physcomitrella DOF genes among the vascular plant sequences belonging to the paraphyletic subfamily A and in a branch preceding that in which subfamilies B–G were anchored, suggests that part of the diversification of the DOF family could have occurred prior to the divergence of the ancestor of mosses and vascular plants. Likewise, the presence of some fern sequences in the subfamily D suggests that this subfamily was formed before the divergence of ferns and the seed plant ancestors. Although the full extent of the diversity of DOF genes in *Pinus taeda* has not been able to be fully established as its genome is not still completely sequenced, the presence of conserved motifs characteristic of subfamilies B, F and G supports that subfamily B originated before the divergence of the angiosperm and the gymnosperm ancestors, and that subfamilies F and G might have diverged in the ancestor of angiosperms after gymnosperm segregation. In addition, the fact that subfamily E was the only one formed by angiosperm genes, indicates that it originated after the divergence of angiosperm and gymnosperm ancestors, whereas subfamily C with only cereal sequences would have been formed after the divergence of the monocot and dicot ancestors.

In conclusion, recurrent duplications of an original DOF formed in the plant ancestor seem to have led to the formation of this complex family of TFs specific to viridiplantae in the course of evolution.

Acknowledgments We thank M. Mena and I. Isabel-LaMoneda for providing the initial λ-based genomic library data. The barley EST EBro02_SQ006_I02 was kindly supplied by Dr R. Waugh and Dr D.F. Marshall, Genome Dynamics/Computational Biology, Scottish Crop Research Institute, Invergowrie (UK). The Institute of Plant Genetics and Crop Plant Research (IPK), Gatersleben (Germany) supplied the following ESTs: HY02C14, HX08G24, HS17N14 and HI14B21. This work was supported by grant no. BMC2003-06345 from Ministerio de Educación y Ciencia (Spain) and by a fellowship (to M.A.M-R.) from Comunidad Autónoma de Madrid (Spain).

References

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

Bailey TL, Elkan C (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc Int Conf Intell Syst Mol Biol 2:28–36

Becker A, Theissen G (2003) The major clades of MADS-box genes and their role in the development and evolution of flowering plants. Mol Phylogenet Evol 29:464–489

Boyes DC, Zayed AM, Ascenzi R, McCaskill AJ, Hoffman NE, Davis KR, Gorlach J (2001) Growth stage-based phenotypic analysis of Arabidopsis: a model for high throughput functional genomics in plants. Plant Cell 13:1499–1510

Dayhoff MO, Schwartz RM, Orcutt BC (1978) A model of evolutionary change in proteins. In: Dayhoff MO (ed) Atlas of protein sequence and structure, vol 5. National Biomedical Research Foundation, Washington, pp 345–352

Doolittle W (1978) Genes in pieces: were they ever together? Nature 271:581–582

Eulgem T, Rushton PJ, Robatzek S, Somssich IE (2000) The WRKY superfamily of plant transcription factors. Trends Plant Sci 5:199–206

Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J (1999) Preservation of duplicate genes by complementary, degenerative mutations. Genetics 151:1531–1545

Gilbert W, de Souza SJ, Long M (1997) Origin of genes. Proc Natl Acad Sci USA 94:7698–7703

Gloeckner G, Beck CF (1995) Genes involved in light control of sexual differentiation in *Chlamydomonas reinhardtii*. Genetics 141:937–943

Gualberti G, Papi M, Bellucci L, Ricci I, Bouchez D, Camilleri C, Costantino P, Vittorioso P (2002) Mutations in the Dof zinc finger genes DAG2 and DAG1 influence with opposite effects the germination of Arabidopsis seeds. Plant Cell 14:1253–1263

Hopfner KP, Kopetzki E, Kresse GB, Bode W, Huber R, Engh RA (1998) New enzyme lineages by subdomain shuffling. Proc Natl Acad Sci USA 95:9813–9818

Huelsenbeck JP, Ronquist FR (2001) MrBayes: Bayesian inference of phylogeny. Bioinformatics 17:754–755

Isabel-LaMoneda I, Diaz I, Martinez M, Mena M, Carbonero P (2003) SAD: a new DOF protein from barley that activates transcription of a cathepsin B-like thiol protease gene in the aleurone of germinating seeds. Plant J 33:329–340

Jakoby M, Weisshaar B, Droge-Laser W, Vicente-Carbajosa J, Tiedemann J, Kroj T, Parcy F (2002) bZIP transcription factors in Arabidopsis. Trends Plant Sci 7:106–111

Joshi CP, Zhou H, Huang X, Chiang VL (1997) Context sequences of translation initiation codon in plants. Plant Mol Biol 35:993–1001

Kettenberger H, Armache KJ, Cramer P (2003) Architecture of the RNA polymerase II-TFIIS complex and implications for mRNA cleavage. Cell 114:347–357

Lijavetzky D, Carbonero P, Vicente-Carbajosa J (2003) Genome-wide comparative phylogenetic analysis of the rice and Arabidopsis Dof gene families. BMC Evol Biol 3:17–28

Long M, Langley CH (1993) Natural selection and the origin of jingwei, a chimeric processed functional gene in Drosophila. Science 260:91–95

Lynch M, Force A (2000) The probability of duplicate gene preservation by subfunctionalization. Genetics 154:459–473

Maizel A, Busch MA, Tanahashi T, Perkovic J, Kato M, Hasebe M, Weigel D (2005) The floral regulator LEAFY evolves by substitutions in the DNA binding domain. Science 308:260–263

Martin C, Paz-Ares J (1997) MYB transcription factors in plants. Trends Genet 13:67–73

Mena M, Vicente-Carbajosa J, Schmidt RJ, Carbonero P (1998) An endosperm-specific DOF protein from barley, highly conserved in wheat, binds to and activates transcription from the prolamin-box of a native B-hordein promoter in barley endosperm. Plant J 16:53–62

Nakai K, Horton P (1999) PSORT: a program for detecting the sorting signals of proteins and predicting their subcellular localization. Trends Biochem Sci 24:34–35

Ohno S (1970) Evolution by gene duplication. Springer, Berlin Heidelberg New York

Papi M, Sabatini S, Bouchez D, Camilleri C, Costantino P, Vittorioso P (2000) Identification and disruption of an Arabidopsis zinc finger gene controlling seed germination. Genes Dev 14:28–33

Parenicova L, de Folter S, Kieffer M, Horner DS, Favalli C, Busscher J, Cook HE, Ingram RM, Kater MM, Davies B, Angenent GC, Colombo L (2003) Molecular and phylogenetic analyses of the complete MADS-box transcription factor family in Arabidopsis: new openings to the MADS world. Plant Cell 15:1538–1551

Pinyopich A, Ditta GS, Savidge B, Liljegren SJ, Baumann E, Wisman E, Yanofsky MF (2003) Assessing the redundancy of MADS-box genes during carpel and ovule development. Nature 424:85–88

Reyes JC, Muro-Pastor MI, Florencio FJ (2004) The GATA family of transcription factors in Arabidopsis and rice. Plant Physiol 134:1718–1732

Roy SW, Gilbert W (2005) Complex early genes. Proc Natl Acad Sci USA 102:1986–1991

Sambrook J, Russell DW (2001) Molecular cloning: a laboratory manual, 3rd edn. Cold Spring Harbor Laboratory Press, New York

Shigyo M, Hasebe M, Ito M (2006) Molecular evolution of the AP2 subfamily. Gene 366:256–265

Tepperman JM, Zhu T, Chang HS, Wang X, Quail PH (2001) Multiple transcription-factor genes are early targets of phytochrome A signalling. Proc Natl Acad Sci USA 98:9437–9442

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin J, Higgins DG (1997) The Clustal X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res 25:4876–4882

Umemura Y, Ishiduka T, Yamamoto R, Esaka M (2004) The Dof domain, a zinc finger DNA-binding domain conserved only in higher plants, truly functions as a Cys2/Cys2 Zn finger domain. Plant J 37:741–749

Vision TJ, Brown DG, Tanksley SD (2000) The origins of genomic duplications in Arabidopsis. Science 290:2114–2117

Yanagisawa S (2002) The Dof family of plant transcription factors. Trends Plant Sci 7:555–560

Yanagisawa S, Schmidt RJ (1999) Diversity and similarity among recognition sequences of Dof transcription factors. Plant J 17:209–214

Yu Y, Tomkins JP, Waugh R, Frisch DA, Kudrna D, Kleinhofs A, Brueggeman RS, Muehlbauer GJ, Wise RP, Wing RA (2000) A bacterial artificial chromosome library for barley (*Hordeum vulgare*) and the identification of clones containing putative resistance genes. Theor Appl Genet 101:1093–1099

Zhang Y, Wang L (2005) The WRKY transcription factor superfamily: its origin in eukaryotes and expansion in plants. BMC Evol Biol 5:1

Zhaxybayeva O, Gogarten JP (2003) Spliceosomal introns: new insights into their evolution. Curr Biol 13:R764–R766

Zimmermann P, Hirsch-Hoffmann M, Henning L, Gruissem W (2004) GENEVESTIGATOR. Arabidopsis Microarray Database and Analysis Toolbox. Plant Physiol 136:2621–2632