J Mol Evol (2003) 56:105–116
DOI: 10.1007/s00239-002-2386-1

JOURNAL OF MOLECULAR EVOLUTION

Evolutionary Pattern of Angiosperm bZIP Factors Homologous to the Maize Opaque2 Regulatory Protein

Michel Vincentz,[1,2] Claudia Bandeira-Kobarg,[1] Luciane Gauer,[1] Paulo Schlögl,[1] Adilson Leite[1]

[1] Centro de Biologia Molecular e Engenharia Genética, Universidade Estadual de Campinas, Cidade Universitaria "Zeferino Vaz," Distrito Barão Geraldo, 13081-970, Campinas, SP, Brazil
[2] Departamento de Genética e Evolução, IB, Universidade Estadual de Campinas, Cidade Universitaria "Zeferino Vaz," Distrito Barão Geraldo, 13081-970, Campinas, SP, Brazil

Received: 3 July 2002 / Accepted: 14 August 2002

Abstract. Opaque2 (O2) is a bZIP transcriptional regulatory factor involved in the control of seed storage proteins synthesis as well as carbon and nitrogen metabolism during maize seed development. Phylogenetic analysis of a possible complete and nonredundant collection of angiosperm bZIP factors resulted in the identification of 20 angiosperm O2-homologues that defined what we call the O2 gene family. Members of the family share a highly conserved bZIP DNA binding domain and several other motifs which define important functional features. The O2 family was enriched by the identification of 25 new putative angiosperm *O2* homologous genes in EST databases and in the rice genome. Based on parsimony analysis, the collection of *O2* homologues was organized into one eudicot–monocot and three monocot groups of orthologous genes and two groups of eudicot genes. These results support a model of the evolution of the O2 family that involves two *O2* homologous gene duplications before the separation of monocots and eudicots. Further expansion of *O2* homologues resulted in at least three and one gene duplications in the monocot and eudicot lineages, respectively. *O2* appears to have been the result of a monocot-specific gene duplication event, and the possibility that *O2* represents a functional specialization restricted to monocots is suggested.

Key words: Angiosperm — *Arabidopsis thaliana* — bZIP factors — Opaque2 — Phylogeny

Introduction

Transcriptional regulatory factors of the basic leucine zipper (bZIP) class have been described in all eukaryotes (Wingender et al. 2000). These factors bind DNA as dimers through a conserved DNA binding domain. This domain is formed by a region rich in basic amino acids that interact with the DNA target site and by a zipper of leucines that consists of several heptad repeats of hydrophobic residues that promote dimerization (Hurst 1995). In angiosperms (flowering plants), Opaque2 (O2) is one of the best-characterized bZIP transcriptional regulators. *O2* is an important regulatory locus of seed endosperm development of the monocot species *Zea mays* (maize) (Schmidt 1993). O2 specifically accumulates in the developing endosperm, where it activates the expression of α- and β-prolamin storage protein genes (Schmidt et al. 1992; Vicente-Carbajosa et al. 1997; Cord Neto et al. 1995), the b-32 albumin genes (Lohmer et al. 1991), and the gene for the cytoplasmic pyruvate orthophosphate dikinase (Gallusci et al. 1996). *O2* is also involved in the control of lysine accumulation (Kemper et al. 1999) and threonine metabolism

Correspondence to: Michel Vincentz; *email:* mgavince@obelix.unicamp.br

(Damerval and Le Guilloux 1998). Overall, these results point to a role for *O2* in the coordinated regulation of storage protein synthesis as well as carbon and nitrogen metabolism during seed development. The recent suggestion that the diurnal modulation of O2 DNA binding activity by phosphorylation/dephosphorylation is regulated by diurnal metabolic fluxes is in line with the latter view (Ciceri et al. 1999).

Identifying O2-related bZIP factors among angiosperms and understanding their evolution are of interest for two reasons: first, because O2 plays an important role in seed development and, second, because O2 has been extensively studied. A phylogenetic analysis using the amino acid sequences of the highly conserved bZIP domain of 50 angiosperm bZIP proteins identified a cluster of 8 monocot and eudicot O2-related proteins, which may form a gene family (Vettore et al. 1998). Although the probable *O2* orthologues from the close relatives of maize, sorghum and *Coix*, are known, the orthologous/paralogous relationships among all the O2-related proteins which are included in this putative family are unclear. More precisely, the existence of *O2* orthologues in eudicotyledonous plants has not been established.

The purpose of this work is to get a detailed picture of the evolution of O2-related proteins in angiosperms, with the additional objective of defining the conditions for a broader analysis of the evolutionary history of angiosperm bZIP factors. To this end, we characterized four cDNAs that represent the complete set of *O2* homologues in the model eudicot plant *Arabidopsis thaliana* (*Arabidopsis*). Additionally, we identified a set of 41 higher-plant *O2* homologous genes, 24 of which were detected in EST databases. This set of genes defines what we call the O2 gene family. Phylogenetic analysis revealed that the evolution of the O2 family could be explained by monocot- and eudicot-specific gene duplication events from three ancestral genes. The possibility that *O2* is restricted to monocot species is also discussed.

Materials and Methods

Cloning of bZIP cDNAs by 3′ Amplification of cDNA Ends, cDNA Library Screening, and DNA Sequencing

The 3′ ends of bZIP cDNAs were amplified in two steps using two pairs of nested degenerated primers and the M13 reverse primer as a 3′-anchor in the cloning vector λ ZAP II. The first pair of nested primers is based on the sequence SNRESARRS, which is conserved in the basic domain of several plant bZIP proteins. The primers are BC5 (5′-TCHAAYMGDGARTCWGC-3′), which corresponds to the peptide SNRESA, and BC6 (5′-aaggaattcGARTCWGCHAGRAGGTC-3′), which corresponds to the peptide ESARRS. The second pair of nested primers is based on the sequence (V/A)KVKM (A/G)E(D/E), which is conserved in the leucine zipper of O2-related proteins. The degenerated primers are ZC3.1 (5′-GYNAAGGTRAAGATGG-3′), which corresponds to the peptide (V/A)KVKM (A/G), and ZC3.2 (5′-aaggaattcGTRAAGATGGSNGARG-3′), which corresponds to the peptide KVKM (A/G)E(D/E). *Eco*RI sites were included in primers BC6 and ZC3.2 to facilitate subsequent cloning.

A cDNA library from *Arabidopsis thaliana* (ecotype Columbia) green siliques, which was constructed in λ ZAP II (Giraudat et al. 1992), was used as the template for the first round of amplification. Approximately 3 · 10^8 phages were amplified with 10 pmol of M13 reverse primer and 150 pmol of either BC5 or ZC3.1 in a 100-μl reaction mixture containing 1.5 m*M* MgCl$_2$, 200 μ*M* deoxynucleotide triphosphate, and 3 U of *Taq* DNA polymerase in the buffer supplied by the manufacturer (BRL). Amplification conditions were 94°C for 4 min followed by 35 cycles (94°C 1 min, 42°C for 2 min, and 72°C for 2 min) and an extension step of 10 min at 72°C. The PCR products were then purified on a QIAquick PCR purification kit (Quiagen) to eliminate unincorporated primers and nucleotides. An aliquot of 3 μl of the purified product was then used in the second round of amplification, which also included 10 pmol of M13 reverse primer and 75 pmol of either primer BC6 or primer ZC3.2. The total reaction volume was 50 μl containing 1.5 m*M* MgCl$_2$, 200 μ*M* deoxynucleotide triphosphate, and 3 U of *Taq* DNA polymerase (BRL). An initial denaturation step at 94°C for 4 min was followed by 36 cycles: 94°C for 1 min, 50°C (first three cycles) and 56°C (the remaining 33 cycles) for 2 min, 72°C for 2 min, and an extension step of 10 min at 72°C. PCR products larger than 200 bp were gel purified, digested with *Eco*RI, and cloned into pBluescript SK$^+$ (Stratagene) for sequencing, which was done by the dideoxy dye terminator method (Perkin Elmer).

Amplification products encoding polypeptides similar to known bZIP proteins were used as probe to isolate the corresponding full-length cDNAs from the seed cDNA library using standard protocols (Sambrook et al. 1989) and starting with 8 · 10^5 recombinant phages. After four plaque purification steps, positive phages were converted to pBluescript SK$^-$ phagemid clones by in vivo excision using the R408 helper phage according to the manufacturer's instructions (Stratagene) except that the last growth step was at 42°C. The complete sequence of the full-length *BZO2H1* and *BZO2H4* (EST T22560; Table 1) cDNAs required the subcloning of an internal *Hin*dIII/*Bam*HI fragment of 700 bp (in pBluescript SK$^+$) and an *Eco*RI fragment of 900 bp (in pUC18), respectively. Standard protocols were used for cloning (Sambrook et al. 1989).

Construction of a Nonredundant Data Set of Angiosperm bZIP Factors and Identification of Tentative Unique Genes from EST Collections

Construction of a nonredundant set of *Arabidopsis* bZIP (Vincentz et al. 2001) and of other angiosperm bZIP proteins was achieved through iterated searches of GenBank protein databases at the National Center for Biotechnology Information (NCBI; http://www.ncbi.nlm.nih.gov) and the MAtDB (Munich Information Center for Protein sequences *Arabidopsis thaliana* Database; http://www.mips.biochem.mpg.de/proj/thal/; v211200) using different known and distantly related bZIP as query sequences and the blastp and tblastn programs (Altschul et al. 1990) at the NCBI (http://www.ncbi.nlm.nih.gov/BLAST/) and MAtDB servers (http://mips.gsf.de/proj/thal/db/search/search_frame.html). Additionally, key word searches were also performed at the NCBI and MAtDB. Editing of exon sequences encoding some of the bZIP domains involved modifications of exon/introns junctions that were guided by amino acid sequence alignments and by the presence of donor (GT) and acceptor (AG) intron splice sites. ESTs related to O2 and its homologous sequences were selected from the GenBank dbEST (http://www.ncbi.nlm.nih.gov/) using tblastn program and protein sequences of O2 homologues (Table 1) as query sequences at the NCBI blast server. Selected ESTs were

Table 1. Opaque2 homologous proteins

	Accession No.		
Protein	Protein	Genomic	cDNA
Eudicot			
BZO2H1 *At*	(N) AAC78255 (M) At4g02640	CRM IV AC002330	AF3 10222
BZO2H2 *At*	(M) At5g24800	CRM V AF069716	AF3 10223
BZO2H3 *At*	(N) AAF67360 (M) At5g28770	CRM V AF262041	AF310224 AA042606
BZO2H4 *At*	(N) CAB77582 (M) At3g54620	CRM III AL138656	AY057509 T22560
BZI-1 *Nt*	AAL27150	—	AY061648
CPRF2 *Pc*	Q99090	—	X58577
Monocot			
O2 *Clj*	S42493	X78287	X78286
BLZ1 *Hv*	T04477	X80068	—
BLZ2 *Hv*	CAA71795	—	Y10834
REB *Os*	BAA36492	ABO21736	—
RISBZ1 *Os*	BAB39173	ABO53475	ABO53472
RISBZ4 *Os*	BAB39174	AAAA1005868[a]	ABO53473
RISBZ5 *Os*	BAB39175	OSJNBb0065C04[b]	ABO53474
RITA1 *Os*	T03990	AJ001267 AAAA01000425[a]	L34551
02 *Sb*	CAA50642	X71636	—
SPA *Ta*	T06767	—	Y09013
02 *Zm*	AAA33489	X15544	M29411
OHP1 *Zm*	JQ2147	—	L00623
OHP1b *Zm*	AAC49533	—	U35063
OHP2 *Zm*	JQ2148	—	L06478

Note. (M) MAtDB Arabidopsis database (http://www.mips.biochem.mpg.de/proj/thal/), (N) NCBI (http://www.ncbi.nlm.nih.gov/). The underlined accession numbers are from the ESTs we refer to in the text. BZO2H1, -H2, and -H3 and the cDNA accession numbers are our own submissions. Species abbreviations: *At*, *Arabidopsis thaliana*; *Clj*, *Coix lacryma-jobi*; *Hv*, *Hordeum vulgare*; *Nt*, *Nicotiana tabacum*; *Os*, *Oryza sativa*; *Pc*, *Petroselinum crispum*; *Sb*, *Sorghum bicolor; Ta*, *Triticum aestivum*; *Zm*, *Zea mays.*

[a] *Oryza sativa* ssp. *indica* (NCBI).

[b] *Oryza sativa* ssp. *japonica* (MATDB).

compared pairwise with the blastn program using default parameters, and overlapping sequences that share at least 98% identity over 150 nucleotides or 100% identity over 100 nucleotides were assembled into a consensus sequence (CS) that defines tentative unique genes (TUGs). Care was taken to avoid probable unspliced variants. A second dbEST search was performed with the TUGs CS to complete our analysis. Finally, an additional assembly step was performed using the PHRAP program (Green 1994). To build a final TUG CS, the CS derived from PHRAP was compared with Clustal X (Thompson et al. 1997) to the TUGs CS obtained with blastn. Rice genomic sequences (ssp. *indica* and *japonicum*) were searched for O2 homologues at the NCBI and MATDB blast servers.

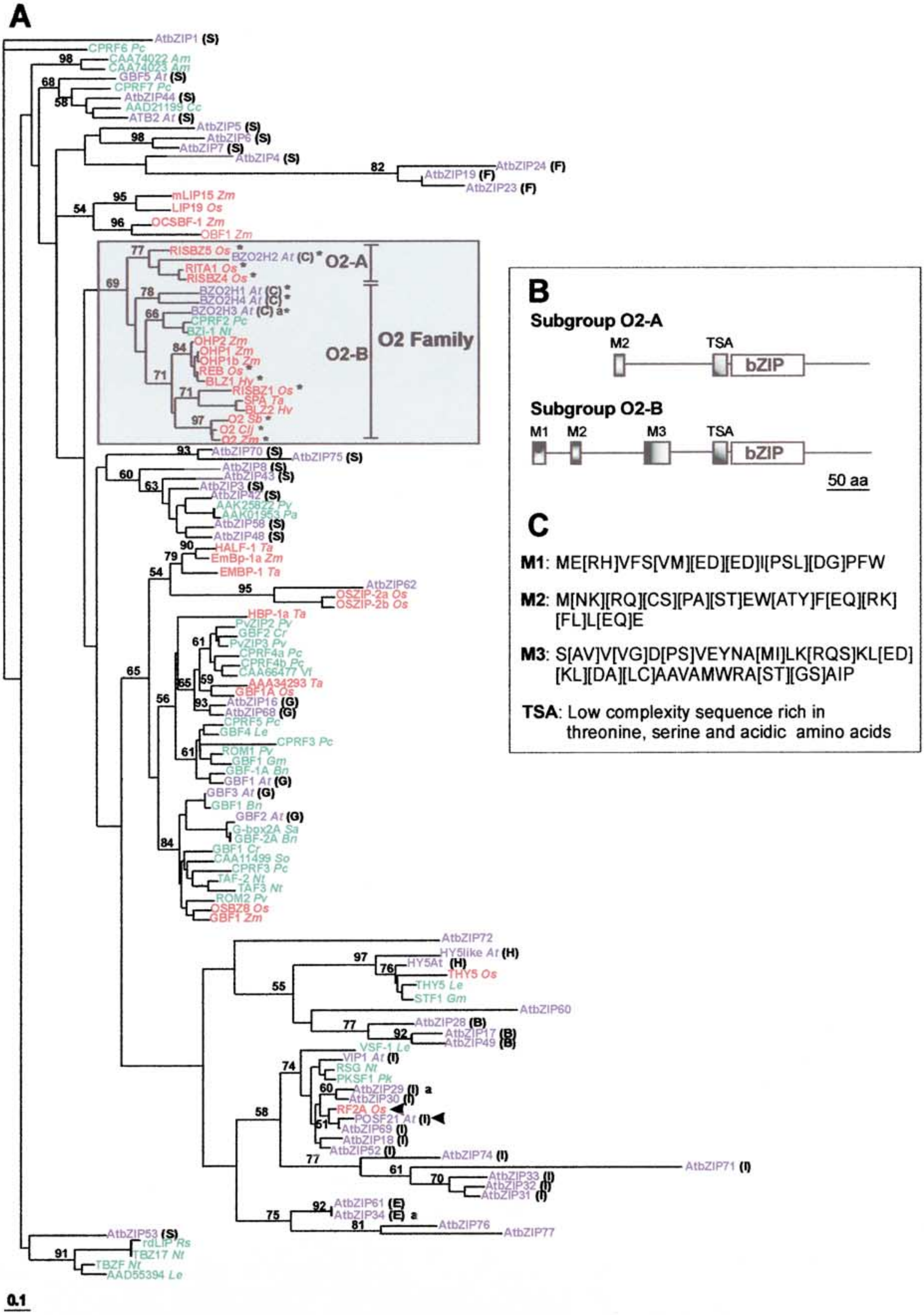
Phylogentic Analysis, Computer Programs, and Web Servers

Comparison of protein sequences with the blastp program (Altschul et al. 1990) was done using default parameters but without filtering. Standard unweighted maximum parsimony analysis was performed on nucleotide sequences by heuristic search using the branch swapping nearest-neighbor interchanges algorithm (search level, 3; initial tree search by random addition option with 100 replications) implemented in the Molecular Evolutionary Genetic Analysis (MEGA) package v 2.1 (Kumar et al. 2000). Neighbor-joining analyses of nucleotide sequence data were performed using different methods of distance estimation that are provided in the program MEGA package v 2.1 (Kumar et al. 2000). Maximum parsimony analysis of amino acid sequences was conducted with the PROTPAR program [PHYLIP, Phylogeny Inference Package version 3.57c (Felsenstein 1993)] and distance method analyses of amino acid sequences data were done with the NEIGHBOR program [PHYLIP, Phylogeny Inference Package version 3.57c (Felsenstein 1993)] using PAM distances (Dayhoff et al. 1978), which were obtained with the PROTDIST program (PHYLIP). The number of synonymous differences per synonymous site (d_s) and the number of nonsynonymous differences per nonsynonymous site (d_n) were calculated based on the method of Pamilo–Biancho–Li (Pamilo and Bianchi 1993; Li 1993) as implemented in MEGA v 2.1. Relative rate test analysis was conducted by the method of Tajima (1993) as implemented in MEGA v 2.1 and using *Arabidopsis BZO2H3* as an outgroup. DNA sequence analysis was carried out with the DNASIS program (Pharmacia). Sequences were aligned with Clustal X (Thompson et al. 1997). Prediction of protein secondary structure was obtained through the predictprotein server (http://www.emblheidelberg.de/predictprotein/) (Rost 1996). Identification of conserved protein motifs was performed with the motif discovering tool MEME (Bailey and Elkan 1994; http://meme.sdsc.edu/meme/website/).

Results

Definition of the Angiosperm O2 Family and of the Modular Protein Structure of Its Members

We identified four *Arabidopsis* cDNAs whose predicted polypeptides were found to be highly similar to O2 based on blastp comparisons. These *Arabidopsis* O2-related proteins were named BZO2H1, BZO2H2, BZO2H3, and BZO2H4 (*b*asic leucine *z*ipper *O2 h*omologous) (Table 1). *BZO2H1* and *BZO2H2* full-length cDNAs were isolated from a seed cDNA library (see Materials and methods). The full-length cDNA sequence of BZO2H3 was reconstructed by joining a partial cDNA that we had isolated from a seed cDNA library and covers the 5′-end mRNA sequence with the overlapping sequence of an EST (accession No. AA042666) that covers the 3′ end of the mRNA sequence. The full-length cDNA sequence of BZO2H4 was obtained from an EST (accession No. T22560). To define the group of monocot and eudicot O2 homologues we performed a phylogenetic analysis of a possibly complete and nonredundant repertoire of 168 *a*ngiosperm *b*Z*I*P factors (ABZ data set) that we constructed from sequences available in public databases. From the complete sequence of the *Arabidopsis* genome (The Arabidopsis Genome Initiative 2000), it was possible to include in the ABZ data set a possible complete and nonredundant set of 76 *Arabidopsis* bZIP proteins. This set of *Arabidopsis*


A
AtbZIP1 (S)
CPRF6 Pc
98
CAA74022 Am
CAA74023 Am
GBF5 At (S)
68
CPRF7 Pc
58
AtbZIP44 (S)
AAD21199 Cc
ATB2 At (S)
AtbZIP5 (S)
98
AtbZIP6 (S)
AtbZIP7 (S)
AtbZIP4 (S)
82
AtbZIP24 (F)
AtbZIP19 (F)
AtbZIP23 (F)
95
mLIP15 Zm
LIP19 Os
54
96
OCSBF-1 Zm
OBF1 Zm
77
RISBZ5 Os *
BZO2H2 At (C) *
O2-A
RITA1 Os *
RISBZ4 Os *
69
78
BZO2H1 At (C) *
BZO2H4 At (C) *
66
BZO2H3 At (C) a*
CPRF2 Pc
BZI-1 Nt
84
OHP2 Zm
OHP1 Zm
OHP1b Zm
O2-B
O2 Family
REB Os *
71
BLZ1 Hv *
71
RISBZ1 Os *
SPA Ta
BLZ2 Hv
97
O2 Sb *
O2 Cij *
O2 Zm *
93
AtbZIP70 (S)
AtbZIP75 (S)
60
AtbZIP8 (S)
AtbZIP43 (S)
63
AtbZIP3 (S)
AtbZIP42 (S)
AAK25822 Pv
AAK01953 Pa
AtbZIP58 (S)
AtbZIP48 (S)
90
HALF-1 Ta
79
EmBp-1a Zm
54
EMBP-1 Ta
AtbZIP62
95
OSZIP-2a Os
OSZIP-2b Os
HBP-1a Ta
PvZIP2 Pv
61
GBF2 Cr
PvZIP3 Pv
CPRF4a Pc
CPRF4b Pc
CAA66477 Vf
65
59
AAA34293 Ta
GBF1A Os
56
93
AtbZIP16 (G)
AtbZIP68 (G)
CPRF5 Pc
GBF4 Le
CPRF3 Pc
61
ROM1 Pv
GBF1 Gm
GBF-1A Bn
GBF1 At (G)
GBF3 At (G)
GBF1 Bn
GBF2 At (G)
G-box2A Sa
GBF-2A Bn
84
GBF1 Cr
CAA11499 So
CPRF3 Pc
TAF-2 Nt
TAF3 Nt
ROM2 Pv
OSBZ8 Os
GBF1 Zm
AtbZIP72
97
HY5like At (H)
76
HY5At (H)
THY5 Os
55
THY5 Le
STF1 Gm
AtbZIP60
77
AtbZIP28 (B)
92
AtbZIP17 (B)
AtbZIP49 (B)
VSF-1 Le
74
VIP1 At (I)
RSG Nt
PKSF1 Pk
60
AtbZIP29 (I) a
AtbZIP30 (I)
RF2A Os
58
POSF21 At (I)
51
AtbZIP69 (I)
AtbZIP18 (I)
AtbZIP52 (I)
77
AtbZIP74 (I)
AtbZIP71 (I)
61
70
AtbZIP33 (I)
AtbZIP32 (I)
AtbZIP31 (I)
92
AtbZIP61 (E)
AtbZIP34 (E) a
75
81
AtbZIP76
AtbZIP77
AtbZIP53 (S)
rdLIP Rs
91
TBZ17 Nt
TBZF Nt
AAD55394 Le
0.1
B
Subgroup O2-A
M2
TSA
bZIP
Subgroup O2-B
M1
M2
M3
TSA
bZIP
50 aa
C
M1: ME[RH]VFS[VM][ED][ED]I[PSL][DG]PFW
M2: M[NK][RQ][CS][PA][ST]EW[ATY]F[EQ][RK][FL]L[EQ]E
M3: S[AV]V[VG]D[PS]VEYNA[MI]LK[RQS]KL[ED][KL][DA][LC]AAVAMWRA[ST][GS]AIP
TSA: Low complexity sequence rich in threonine, serine and acidic amino acids

bZIP factors was obtained by the integration of our own data (Vincentz et al. 2001) and those of The bZIP Research Group (Jakoby et al. 2002) and is slightly smaller than the set of 81 bZIP proteins described earlier (Riechmann et al. 2000).

In agreement with previous results (Vettore et al. 1998), significant amino acid sequence similarity among all these angiosperm bZIP proteins was found to be restricted to the minimum bZIP DNA binding domain, which consists of the basic motif and three leucine repeats (of 44 amino acids, which correspond to positions 228 to 271 in the maize O2 sequence: Fig. 2). A neighbor-joining analysis of the minimum bZIP domain (amino acid sequences) of the ABZ data set revealed the existence of two large clusters of proteins with significant bootstrap support (results not shown). These two clusters include all *Arabidopsis* bZIP factors of family III/group A (51% bootstrap support) and family II/group D (98% bootstrap support) of Vincentz et al. (2001) and Jakoby et al. (2002), respectively. As these two clusters did not contain O2 and were found to be responsible for the phylogenetic analysis of the ABZ data set being restricted to the minimal bZIP domain, their members were excluded from the ABZ collection to create a subset of the ABZ collection (SABZ data set) that includes 122 bZIP proteins. The length of the sequences that could be aligned from the SABZ data set was increased by two leucine repeats (16 amino acids) compared to the minimal bZIP domain. This, in turn, allowed an improvement in the resolution of the evolutionary relationships among the SABZ set of proteins. The unrooted tree inferred from a neighbor-joining analysis of the bZIP domain amino acid sequences of the SABZ data set is shown in Fig. 1A. This tree identifies a well-supported group of proteins, which most likely are O2 homologues, and was defined as the O2 family (Fig. 1A). This family is formed by 6 eudicot (including the complete set of *Arabidopsis* O2 homologues BZO2H1, BZO2H2, BZO2H3, and BZO2H4) and 14 monocot proteins (Table 1). The gene structure of 13 members of the O2 family is available and was compared. For all of them, 78% of the bZIP domain is encoded by exons 4 and 5, whose size and positions are conserved (Fig. 2). Furthermore, this feature appears to be specific to members of the O2 family as judged by the analysis of the complete set of *Arabidopsis* genes encoding bZIP factors (data not shown). These results indicate that this set of 13 genes of the O2 family are homologous and support the notion that the set of 20 angiosperm bZIP factors that are included in the O2 family are indeed O2 homologous proteins.

The proteins of the O2 family share a highly conserved bZIP domain which has the double function of DNA binding and nuclear localization (Varagona and Raikel 1994). Prediction of coiled-coil structures by the Coil program (Lupas 1996) indicated that up to nine leucine (hydrophobic residues) heptad repeats may be involved in the leucine zipper dimerization domain of the O2-related proteins (Fig. 2). Additionally, a set of conserved motifs which may define important functional sequences was identified (Fig. 1B). Motifs M1, M2, and TSA participate in

◀

Fig. 1. **A** Phylogenetic tree defining the O2 family. The unrooted tree was inferred by the neighbor-joining method using PAM distances of bZIP-domain amino acid sequences (basic motif plus five leucines of the leucine zipper; positions 228 to 287 of the maize O2 protein in Fig. 2) of approximately two-thirds of the complete and nonredundant set of angiosperm bZIP factors that are available in databases. Bootstrap values of 500 replicates are indicated as percentages along the branches. The O2 family is *boxed in gray*. Known gene structures of members of the O2 family are marked with an *asterisk*. *Arabidopsis*, other eudicot, and monocot proteins are shown in *blue*, *green*, and *red*, respectively. The bZIP-domain amino acid sequences of proteins marked with an *a* were edited (see Materials and Methods). AtbZIP76 (accession No. AAG50695) and AtbZIP77 (accession No. NP_564460) are putative new *Arabidopsis* proteins. AtbZIP73 (MATDB accession No. At2g13130) presents a stop codon in the leucine zipper coding part and was therefore not included in our data set. The classification of *Arabidopsis* bZIPs established by The bZIP Research Group (Jakoby et al. 2002) is indicated in *parentheses* (B, C, E, F, G, H, I, and S). The two *arrows* point to the outgroup used in the parsimony analysis in Fig. 3. Species abbreviations and accession numbers for the O2 family members are as in Table 1. Accession numbers of *Arabidopsis* proteins are those of The bZIP Research Group (Jakoby et al. 2002). Other accession numbers for eudicot proteins are as follows: rdLIP *Rs*, BAA34938; TBZF *Nt*, BAB13719; CPRF6 *Pc*, CAC00657; CPRF7 *Pc*, CAC00658; TBZ17 *Nt*, BAA22204; GBF1 *Cr*, AAD42937; GBF2 *Cr*, AAD42938; PvZIP2 *Pv*, AAK39130; PvZIp3 *Pv*, AAK39131; CPRF5 *Pc*, CAC00656; GBF1 *Gm*, AAB00096; G-box2A *Sa*, T10472; GBF-1A *Bn*, CAA58774; GBFl *Bn*, AAB03379; GBF-2A *Bn*, CAA58772; ROM2 *Pv*, AAC49474; ROM1 *Pv*, AAB36514; CPRF3 *Pc*, CAA41452; GBF4 *Le*, CAA52896; CPRF4a *Pc*, CAA71768; CPRF4b *Pc*, CAA71770; TAF-2 *Nt*, CAA88492; TAF-3 *Nt*, CAA88493; THY5 *Le*, CAB57979; STF1 *Gm*, AAC05017; RSG *Nt*, BAA97100; PKSFl *pk*, AAC04862; VSF-1 *Le*, CAA05898. For monocot proteins accession numbers are as follows: mLIP15 *Zm*, BAA05117; Lip19 *Os*, CAA40596; OCSBF-1 *Zm*, CAA44607; OBF1 *Zm*, JQ0984; HBP-1a *Ta*, BAA02304; HALF-1 *Ta*, BAA10928; EMBP-1 *Ta*, AAA68428; GBF1A *Os*, T03241; OSBZ8 *Os*, AAB40291; OSZIP2a *Os*, AAC49557; OSZIP2b *Os*, AAC49558; GBF1 *Zm*, AAA80169; EMBP-la *Zm*, CAB62402; THY5 *Os*, BAB62558; RF2a *Os*, AAC49832. Species abbreviations not listed in Table 1 are as follows: *Am*, *Antirrhinum majus*; *Bn*, *Brassica napus*; *Cc*, *Capsicum chinense*; *Cr*, *Catharanthus roseus*; *Gm*, *Glycine max*; *Le*, *Lycopersicon esculentum*; *Pa*, *Phaseolus acutifolius*; *Pk*, *Paulownia kawakamii*; *Pv*, *Phaseolus vulgaris*; *Rs*, *Raphanus sativus*; *Sa*, *Sinapsis alba*; *So*, *Spinacia oleracea*; *Vf*, *Vicia faba*. The scale bar corresponds to 0.1 estimated amino acid substitution per site. **B** Modular organization of O2 homologous proteins. The distribution of the conserved motifs M1, M2, M3, and TSA along the protein sequence is schematized. These motifs were defined with the help of the MEME tool (Bailey and Elkan 1994). bZIP, basic leucine zipper DNA binding domain shown in Fig. 2. **C** Multilevel consensus sequences (as defined by MEME) of the conserved motifs shown in B.

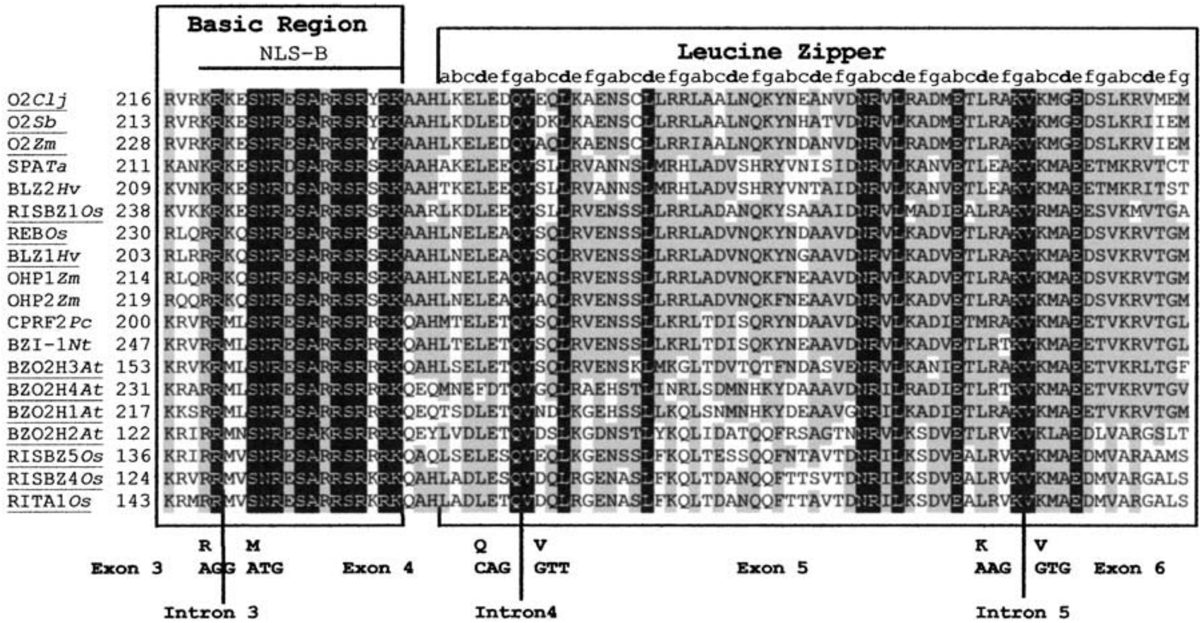

Basic Region
NLS-B
Leucine Zipper
abcdefgabcdefgabcdefgabcdefgabcdefgabcdefgabcdefgabcdefgabcdefg
O2Clj 216 RVRKRKESNRESARRSRYRKAAHLKELEDQVEQLKAENSCLLRRLAALNQKYNEANVDNRVLRADMETLRAKVKMGEDSLKRVMEM
O2Sb 213 RVRKRKESNRESARRSRYRKAAHLKDLEDQVDKLKAENSCLLRRLAALNQKYNHATVDNRVLKADMETLRAKVKMGEDSLKRIIEM
O2Zm 228 RVRKRKESNRESARRSRYRKAAHLKELEDQVAQLKAENSCLLRRIAALNQKYNDANVDNRVLRADMETLRAKVKMGEDSLKRVIEM
SPATa 211 KANKRKESNRDSARRSRSRKAAHAKELEEQVSLLRVANNSLMRHLADVSHRYVNISIDNRVLKANVETLEAKVKMAEETMKRVTCT
BLZ2Hv 209 KVNKRKESNRDSARRSRSRKAAHTKELEEQVSLLRVANNSLMRHLADVSHRYVNTAIDNRVLKANVETLEAKVKMAEETMKRITST
RISBZ1Os 238 KVKKRKESNRESARRSRSRKAARLKDLEEQVSLLRVENSSLLRRLADANQKYSAAAIDNRVLMADIEALRAKVRMAEESVKMVTGA
REBOs 230 RLQRRKQSNRESARRSRSRKAAHLNELEAQVSQLRVENSSLLRRLADVNQKYNDAAVDNRVLKADVETLRAKVKMAEDSVKRVTGM
BLZ1Hv 203 RLRRRKQSNRESARRSRSRKAAHLNELEAQVSQLRVENSSLLRRLADVNQKYNGAAVDNRVLKADVETLRAKVKMAEDSVKRVTGM
OHP1Zm 214 RLQRRKQSNRESARRSRSRKAAHLNELEAQVAQLRVENSSLLRRLADVNQKFNEAAVDNRVLKADVETLRAKVKMAEDSVKRVTGM
OHP2Zm 219 RQQRRKQSNRESARRSRSRKAAHLNELEAQVAQLRVENSSLLRRLADVNQKFNEAAVDNRVLKADVETLRAKVKMAEDSVKRVTGM
CPRF2Pc 200 KRVRRMLSNRESARRSRRRKQAHMTELETQVSQLRVENSSLLKRLTDISQRYNDAAVDNRVLKADIETMRAKVKMAEETVKRVTGL
BZI-1Nt 247 KRVRRMLSNRESARRSRRRKQAHLTELETQVSQLRVENSSLLKRLTDISQKYNEAAVDNRVLKADVETLRTKVKMAEETVKRVTGL
BZO2H3At 153 KRVKRMLSNRESARRSRRRKQAHLSELETQVSQLRVENSKLMKGLTDVTQTFNDASVENRVLKANIETLRAKVKMAEETVKRLTGF
BZO2H4At 231 KRARRMLSNRESARRSRRRKQEQMNEFDTQVGQLRAEHSTLINRLSDMNHKYDAAAVDNRILRADIETLRTKVKMAEETVKRVTGV
BZO2H1At 217 KKSRRMLSNRESARRSRRRKQEQTSDLETQVNDLKGEHSSLLKQLSNMNHKYDEAAVGNRILKADIETLRAKVKMAEETVKRVTGM
BZO2H2At 122 KRIRRMNSNRESAKRSRRRKQEYLVDLETQVDSLKGDNSTLYKQLIDATQQFRSAGTNNRVLKSDVETLRVKVKLAEDLVARGSLT
RISBZ5Os 136 KRIRRMVSNRESARRSRRRKQAQLSELESQVEQLKGENSSLFKQLTESSQQFNTAVTDNRILKSDVEALRVKVKMAEDMVARAAMS
RISBZ4Os 124 KRVRRMVSNRESARRSRKRKQAHLADLESQVDQLRGENASLFKQLTDANQQFTTSVTDNRILKSDVEALRVKVKMAEDMVARGALS
RITA1Os 143 KRMRRMVSNRESARRSRKRKQAHLADLETQVDQLRGENASLFKQLTDANQQFTTAVTDNRILKSDVEALRVKVKMAEDMVARGALS
R
M
Exon 3
AGG
ATG
Exon 4
Q
CAG
V
GTT
Exon 5
K
AAG
V
GTG
Exon 6
Intron 3
Intron4
Intron 5

Fig. 2. Alignment of bZIP amino acid sequences of angiosperm O2 homologous proteins. The basic region includes a nuclear localization signal (NLS). The position of the amino acids in each heptad repeat of the leucine zipper is indicated (a to g) and the position (d) of leucines (hydrophobic residues) is shown in **bold face**. Identical and conserved amino acids are *boxed in gray* when present in at least 50% of the proteins. Strictly conserved residues are highlighted in *black*. Proteins whose gene structures are known are *underlined* and the positions of introns 3, 4, and 5 are shown. Species abbreviations and accession numbers are as in Table 1. The sequence of OHP1b is 97% identical to OHP1 and was therefore not included.

transcriptional activation (Schmitz et al. 1997; Vincente-Carbajosa et al. 1998; Onodera et al. 2001), and the control of nucleocytoplasmic distribution partly involves the TSA and M3 motifs (Varagona and Raikhel 1994; Kircher et al. 1999).

These motifs can be considered as informative shared derived characters that were used to classify members of the O2 family further into the two subgroups (clades) O2-A and O2-B, which are distinguished by the presence/absence of motifs M2 and M3 (Fig. 1B). This classification is consistent with the phylogenetic analysis where members of subgroup O2-A form a well-supported cluster of proteins (Fig. 1A).

Evolution of the O2 Gene Family

Our main objective in initiating a detailed analysis of the evolution of the O2 family was to establish orthologous relationships among members of the family through the identification of *g*roups of *o*rthologous genes (GO). A GO consists of individual orthologous genes or orthologous groups of paralogues from several lineages (Tatusov et al. 1997). Orthologous and paralogous genes are homologous genes that result from a speciation event and from a duplication event within a lineage, respectively (Tatusov et al. 1997; Fitch 2000; Thornton and DeSalle 2000). An important aspect of defining GOs is that it should facilitate the identification of ancestral genes and should be useful to rationalize the systematic analysis of yet uncharacterized proteins (Thornton and DeSalle 2000).

In the first step, to improve the phylogenetic analysis of the angiosperm O2 family, we searched for new O2 homologues that would be represented in plant EST databases. ESTs selected through iterated searches in the GenBank plant dbEST were assembled to form a consensus sequence (CS) that represents tentative unique genes (TUGs). The polypeptides corresponding to each TUG were confirmed to belong to the O2 family if they had the best blastp match with one of the O2 homologues defined here (Table 1). Based on the 96.5% amino acid identity that was observed between the coding sequences of the two maize recent paralogues, OHP1 and OHP1b (Pysh and Schmidt, 1996), only those CSs whose deduced amino acid sequence showed less than 97% identity over their full-length sequence with one of the O2 homologues were considered to represent new genes. Our scheme resulted in the identification of 11 monocot and 13 eudicot putative new *O2* homologous genes. We noticed, however, that in some cases, nonoverlapping CSs whose deduced polypeptides show a higher similarity to the same O2 homologue might actually identify the same gene. Additionally, we identified in the recently published rice genome (Yu et al. 2002) a new *O2* homologous gene (*RBZO2H*). This new *O2* homologous rice gene and nine of the TUGs which

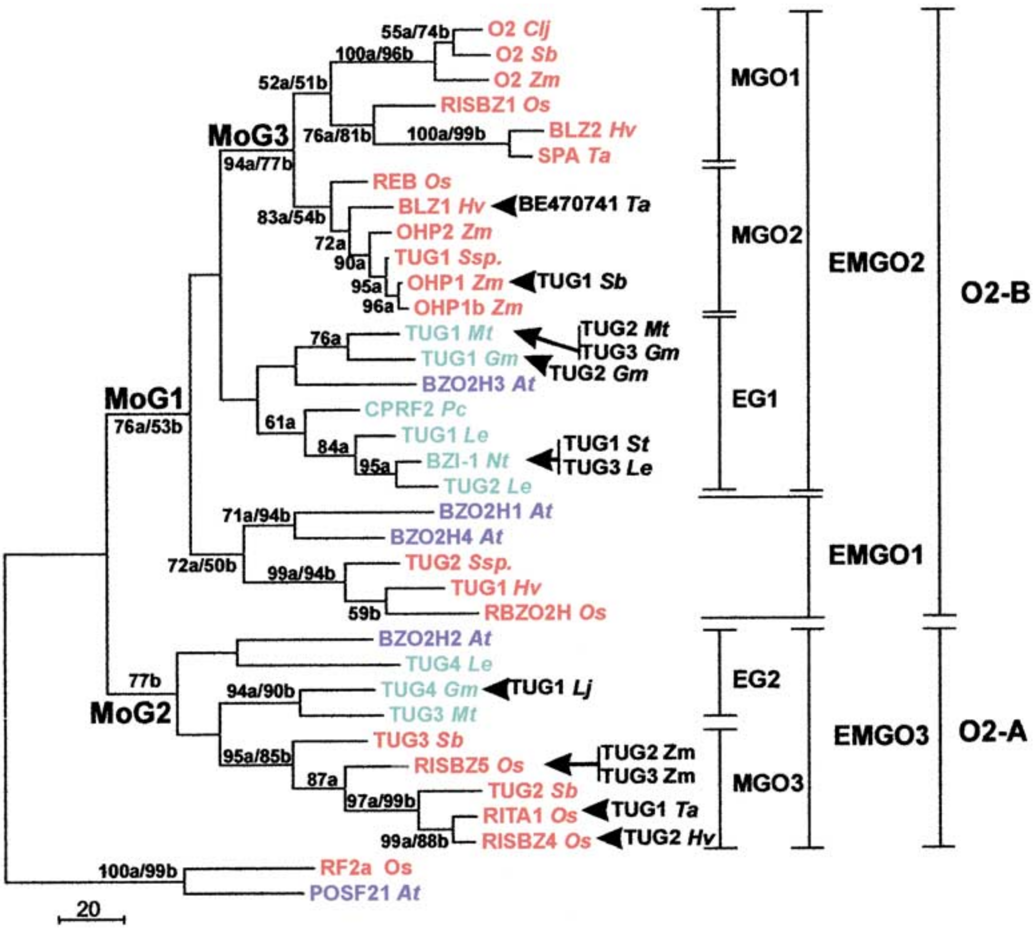


Fig. 3. Phylogeny of the *O2* homologous genes. A representative maximum parsimony rooted tree is shown. Parsimony analysis of bZIP-domain (positions 228 to 313 of maize O2; Fig. 2) DNA sequences of *O2* homologues was obtained by the nearest-neighbor interchange (branch swapping) algorithm included in the MEGA 2 package v 2.1 (Kumar et al. 2000). Bootstrap support over 50% obtained from an analysis including all three sites of each codon (a) or obtained from an analysis including the first two nucleotides of each codon (b) are shown along the branches. For the analysis with three sites, the number of informative sites was 204/258; the tree length, 1241; the informative site consistency index, 0.36; the informative site retention index, 0.6; and the informative site rescaled consistency index, 0.22. For the analysis with two sites, the number of informative sites was 119/172; the tree length, 477; the informative site consistency index, 0.49; the informative site retention index, 0.72; and the informative site rescaled consistency index, 0.35. *Arabidopsis*, other eudicot, and monocot proteins are shown in *blue*, *green*, and *red*, respectively. The GenBank accession number of the genomic sequence (*Oryza sativa* ssp. *indica*) encoding RBZO2H is AAAA01005225. Tentative unique genes (TUGs) that were not included in the phylogenetic analysis point to their closest related O2 homologue in the tree (*arrow*). Highest similarity of the TUGs that were not included in the parsimony analysis to one of the O2 homologues was defined by protein distances that were measured with the PRODIST program (PHYLIP) on aligned amino acid sequences that were obtained with Clustal X using default parameters. MoG, *m*onophyletic *g*roup of proteins; MGO, monocot group of orthologues; EG, eudicot group of genes; EMGO, eudicot–monocot group of orthologues. O2-A and O2-B refer to the two subgroups defined in Fig. 1. Species abbreviations are as in the legend to Fig. 1, and *Ssp.* stands for *Saccharum* sp. The scale bar represents 20 substitutions (obtained from an analysis including the three nucleotides of codons).

encoded a complete bZIP domain were included in the phylogenetic analysis of the O2 family presented hereafter.

The evolutionary history of the O2 family was estimated from the bZIP DNA binding domain, which is the unique well-defined structural and functional domain present in all O2 homologues. The phylogeny of the O2 family was evaluated by a maximum parsimony analysis of nucleotide sequences of the bZIP domain shown in Fig. 2. This analysis included a pair of outgroup sequences formed by the monocot RF2a gene from rice and by the POSF21 gene from *Arabidopsis* (Fig. 1). These sequences were chosen as outgroups based on the shared position of an intron in the bZIP domain with *O2* homologues, indicating that these outgroups and the O2 family most likely derive from a common ancestor (result not shown). The rooted tree inferred from the parsimony analysis is shown in Fig. 3. Parsimony analysis of bZIP domain amino acid sequences as well as distance methods (neighbor joining) applied to nucleotide and amino acids sequences of the bZIP domain data set gave essentially the same results (data not shown).

To interpret the tree shown in Fig. 3, we considered the following three criteria. First, bootstrap support over 50% was retained for the branching pattern. Second, assuming that the complete set of *Arabidopsis* and rice *O2* homologues was identified and no selective gene loss occurred, each group of eudicot orthologues should include at least one *Arabidopsis* gene, each group of monocot orthologues should include at least one rice gene, and each group of eudicot–monocot orthologues should include at least one *Arabidopsis* and one rice gene. Third, the inferred gene phylogeny should be consistent with the known species phylogeny. Accordingly, the tree in Fig. 3 was organized into two main *mo*nophyletic groups, MoG1 and MoG2. The consistency between the composition of these two monophyletic groups and the composition of the two subgroups O2-A and O2-B, which were defined by different means (Fig. 1), provides additional support for the definition of MoGO1 and MoG2. MoG1 was further divided into MoG3, which is formed by the two *m*onocot *g*roups of *o*rthologous genes MGO1 (includes O2) and MGO2, the *e*udicot *g*roup of proteins EG1, and the *e*udicot–*m*onocot *g*roup of *o*rthologues EMGO1 (Fig. 3). MoG2 was resolved into the eudicot group of genes EG2 and the monocot group of orthologues MGO3 (Fig. 3). This general organization was supported by the observation that the polypeptides encoded by the six monocot and six eudicot TUGs, which were not included in our phylogenetic analysis because they do not cover a complete bZIP domain, are more closely related to the monocot and eudicot O2 homologues, respectively (Fig. 3). The relationships among the eudicot genes in EG1 and EG2 (Fig. 3) could not be resolved clearly, essentially because the two *Arabidopsis* genes *BZO2H3* (EG1) and *BZO2H2* (EG2) do not cluster significantly with any of the other eudicot genes of these groups. However, considering that EG1 and EG2 fulfill the second and third criteria used earlier to interpret our tree (see above), they are likely to represent eudicot groups of orthologues. We also noticed that the two *Arabidopsis* genes in EMGO1, *BZO2H1* on chromosome IV and *BZO2H4* on chromosome III, are part of two conserved and collinear segments formed by two matching genes (result not shown). These data indicate that *BZO2H1* and *BZO2H4* are two paralogues that probably arose with one of the large-scale duplications that formed the *Arabidopsis* genome (Vision et al. 2001). Finally, the orthologous relationship among members of MGO1, which includes O2 (Fig. 3), was further supported by their shared seed-specific expression, which is a characteristic restricted to this group of genes (Schmidt et al. 1992; Albani et al. 1997; Yunes et al. 1998; Oñate et al. 1999; Onodera et al. 2001; our unpublished data). Together these data suggest that the O2 family can be organized into three eudicot–monocot groups of orthologous genes (EMGO1, -2, and -3; Fig. 3).

Table 2. d_s and d_n among members of the monocot group of orthologues MGO1 and monocot group of orthologues MGO2

	d_s ($\pm$SE)	d_n ($\pm$SE)
MGO1: *O2 Zm–O2 Sb*	0.408 (0.108)	0.049 (0.017)
MGO2: OHP1 *Zm*–TUG2 *Sb*	0.086 (0.039)	0

Note. d_s and d_n were estimated over 234 nucleotides that encode 91% of the bZIP domain (positions 236 to 313 for the maize O2 factor: Fig. 2). Species abbreviations are as in Table 1, Note.

Our phylogenetic analysis also indicates that the monocot MGO1 and MGO2 emerged as the result of a gene duplication that must have occurred after the separation of monocots and eudicots and some time before the radiation of grasses (Fig. 3). To characterize this duplication event further, we decided to define d_s and d_n for pairs of orthologues in each of these two MGOs. To obtain reliable estimations for d_s ($d_s < 0.5$), we used the maize and sorghum *O2* sequences in MGO1 and their corresponding paralogues, the maize *OHP1* and the sorghum *TUG1* sequences in MGO2 (Fig. 3). We also verified the homogeneity of substitution rates among these two groups of orthologues by applying Tajima's (1993) relative rate test. No rate heterogeneity was observed among the maize and the sorghum *O2* (MGO1) or among the maize *OHP1* and the sorghum *TUG2* orthologues (MGO2). However, these two pairs of orthologues evolved at a significantly different rate (result not shown). Estimations of d_s and d_n were then calculated for 91% of the bZIP sequence, which is the limit imposed by the sorghum TUG2. As shown in Table 2, d_s and d_n were higher in MGO1 than in MGO2. The same trend was observed for the rice *RISBZ1* and the barley *BLZ2* genes (MGO1) and the corresponding paralogues in MGO2, the rice *REB* and the barley *BLZ1* genes (result not shown). Our data suggest that genes in MGO1 (*O2* orthologues) evolved more rapidly than their corresponding paralogues in MGO2.

Discussion

As the first step toward a broad analysis of bZIP factor evolution in angiosperms, we choose to focus our analysis on the origin of the maize *O2* regulatory locus, which is one of the best-characterized bZIP proteins (see Introduction). Additionally, our interest was to identify *O2* orthologues in eudicot species. To this end, we developed a two-step phylogenetic approach. First, from the analysis of the possible complete and nonredundant set of known angiosperm bZIP proteins (ABZ data set), we defined a subset of bZIP factors (SABZ data set) that allowed us, in the second step, to identify a group of 20 O2 homologous proteins (the O2 family; Fig. 1). Our analysis also allowed us to identify several other clusters of proteins with significant bootstrap support (Fig. 1) and this grouping was found to be consistent with classification schemes described earlier for the *Arabidopsis* bZIP proteins (Vincentz et al. 2001; Jakoby et al. 2002). Taken together, our results indicate that the ABZ and SABZ data sets should be useful to improve our knowledge about the evolution of angiosperm bZIP factors.

We further characterized the *O2* homologous genes by determining the modular structure of the corresponding proteins. The functionally essential bZIP DNA binding domain is highly conserved and includes a leucine zipper that is possibly formed by nine leucines (hydrophobic residues). Conserved motifs involved in the control of nuclear translocation or transcriptional activation were also identified (Fig. 1B). Such functional motifs can be considered as shared derived characters and allowed us to divide the O2 family into the two evolutionary distinct subgroups O2-A and O2-B (Fig. 1). Finally, we notice that more than half of the protein sequence of all O2 homologues is poorly conserved, which could reflect either weak functional constraints or functional diversification as has been suggested in the case of the *R* family of basic helix–loop–helix regulatory genes (Purugganan and Wessler 1994). These observations indicate that an accurate phylogenetic analysis of the O2 family should rely mainly on the sequence of the bZIP domain, and this seems to be true for the majority of the others bZIP families (Fig. 1 and unpublished results).

The interpretation of the parsimony analysis of the bZIP domain of O2 homologues relied on (1) bootstrap support over 50%, (2) the assumption that the complete set of *Arabidopsis* and rice O2 homologues was identified, and (3) congruence between the inferred gene phylogeny and the known species phylogeny. Following these criteria, the examination of the tree inferred from parsimony analysis led to the identification of three eudicot/monocot groups of orthologues (EMGO1, -2, and -3 in Fig. 3). An effort was made to incorporate information extracted from EST databases into the phylogenetic analysis. This approach was shown here to improve the reliability of the organization of homologous genes into groups of orthologues.

The monocot species in our analysis are represented exclusively by members of the grass family, which diverged approximately 60 million years ago (MYA) (Kellog 2001). Identifying groups of orthologues among grasses should therefore be facilitated by the fact that they form a compact group of species.

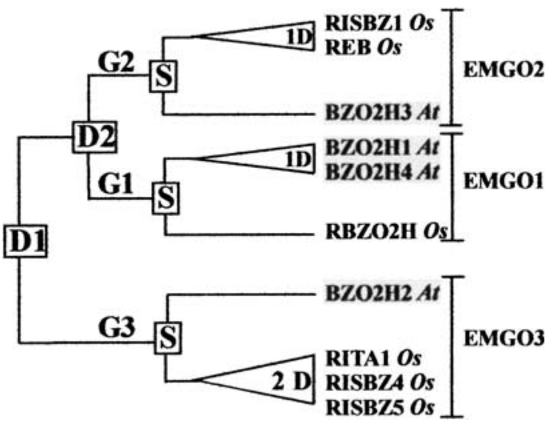


Fig. 4. Models of O2 family evolution. In the ancestral lineage of angiosperms, two duplications, D1 and D2, produced the three *O2* homologues G1, G2, and G3 (ancestral functions). Following the separation (S) of monocots and eudicots, lineage-specific duplications (D) occurred. Representative eudicot genes were from *Arabidopsis* (*At*) and are *boxed in gray*, and representative monocot genes are from rice (*Os*).

However, whenever groups of orthologues evolve at different rates, such as in the case of the MGO1 and MGO2 genes (Fig. 3), additional information such as expression pattern, functional hints, or map position may be required to support orthology. On the other hand, the difficulty of establishing the relationships among eudicot genes, as in the case of the genes included in EG1 and EG2 (Fig. 3), is due partly to the early radiation of eudicot species, such as tomato (*Lycopersicon esculentum*) and *Arabidopsis*, which diverged about 112–156 MYA (Yang et al. 1999). The general strategy presented here sets the conditions for a broader analysis of the evolutionary relationship among angiosperm bZIP factors and the recent publication of the rice genome (Goff et al. 2002; Yu et al. 2002) should significantly improve such an analysis.

The phylogenetic analysis of the O2 family (Fig. 3) suggests a single model to explain the evolutionary history of the angiosperm O2 homologues (Fig. 4). In this model, three O2 homologous genes were present in the angiosperm ancestral lineage. After the separation of eudicots and monocots, lineage-specific gene duplications further shaped the O2 family. This model underscores the potential for functional diversification or innovation from the lineage-specific gene duplication events (Ohno 1970; Walsh 1995; Hughes 1994; Nadeau and Sankoff 1997; Zhang et al. 1998; Gu 1999; Lynch and Conery 2000; Lynch and Force 2000; Ohta 2000; Wendel 2000). For instance, the monocot–dicot EMGO3 (Fig. 3) is formed by one eudicot gene, which is likely to represent an ancestral function, and up to three monocot genes, which have diverged from each other and consequently provided the opportunity to acquire new functional specificity.

Not much is known about the possible functions of the factors included in EMGO3 (but see Onodera et al. 2001). Reverse genetic approaches aimed at inactivating the unique *Arabidopsis* gene *BZO2H2* that is included in EMGO3 is an obvious approach to learn more about EMGO3 genes.

A more informative example is provided by MGO1 (O2 orthologues) and MGO2. These two MGOs were produced by a gene duplication that is restricted to monocots and that must have happened some time before the diversification of grasses (Fig. 3). The large differences in d_s and d_n between MGO1 and MGO2 genes (Table 2) most likely reflects altered functional constraints between these two MGOs and supports functional change between them (Zhang et al. 1998; Gu 1999; Graur and Li 2000). As this conclusion is based on the analysis of the bZIP domain, functional divergence between MGO1 and MGO2 genes may concern DNA binding and/or dimerization specificity. Functional divergence between MGO1 and MGO2 genes is further supported by molecular data gained from the maize *O2* gene (MGO1) and its paralogue *OHP1* (MGO2). For instance, the *in vitro* interaction of O2 with a member of the Dof class of plant Cys_2–Cys_2 zinc-finger DNA binding protein is not shared by OHP1 (Vicente-Carbajosa et al. 1997). Additionally, the expression patterns of *O2* and *OHP1* are different. O2 is specifically expressed in the endosperm and is under the control of a circadian clock, while OHP1 is expressed in the endosperm, root, shoots, leaves, and embryos and is not regulated by an endogenous clock (Ciceri et al. 1999; Pysh et al. 1993). Changes in the coding and regulatory sequences seem, therefore, to have contributed to the functional divergence between MGO1 and MGO2 genes (Wendel 2000). *O2* orthologues (MGO1) may have evolved toward the function of integrating the synthesis of prolamine seed storage proteins with carbon and nitrogen metabolism during endosperm development. The fact that those prolamines are considered to have appeared specifically in grasses (Shewry et al. 1995) raises the interesting possibility of a coordinated evolution of prolamines and *O2*. An implicit consequence of this model is that *O2* is a function that has been recently acquired, while MGO2 genes, which seem to be under stronger selective constraints (Table 2), represent a conserved ancestral function possibly involved in the control of some aspect of carbon and nitrogen metabolism. One predominant adaptive event in this model may have been the acquisition of endosperm-specific expression by *O2*. It remains to be determined if neofunctionalization (Walsh 1995) or subfunctionalization (Lynch and Force 2000) is responsible for the preservation of MGO1 and MGO2 genes.

O2 controls lysine degradation during maize endosperm development by regulating the expression of the gene encoding lysine–ketoglutarate reductase/saccharopine deshydrogenase (*LKR/SDH*) (Kemper et al. 1999), and it was recently shown that knockout *Arabidopsis LKR/SDH* mutants accumulate lysine during seed development (Zhu et al. 2001). It appears, therefore, that control of lysine catabolism as a means of regulating its accumulation in seeds is conserved among angiosperms. It will be interesting to see if the *Arabidopsis* bZIP factor BZO2H3, which is more closely related to the monocot MGO1 and MGO2 genes (Fig. 3), is involved in this regulatory process.

Acknowledgments. The authors thanks Dr. J. Giraudat (Institut des Sciences Végétales, Gyf sur Yvette) for the *Arabidopsis thaliana* cDNA library and the Arabidopsis DNA Stock Center for ESTs. We also acknowledge the support of Dr. P. Arruda and L. Grivet, F.R. Silva, and L.B. Klaczko for critically reading the manuscript. S.L. Martins Silva is acknowledged for technical assistance. This work was supported by a grant from Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP)/Projeto Temático de Equipe (PTE)-99/02198-3.

References

Albani D, Hammond-Kosack MCU, Smith C, Conlan S, Colot V, Holdsworth M, Bevan M (1997) The wheat transcriptional activator SPA: A seed specific bZIP protein that recognizes the GCN4-like motif in the bifactorial endosperm box of prolamin genes. Plant Cell 9:171–184

Altschul SF, Gish W, Miller W, Myers EW, Lipman D (1990) Basic local alignment tool. J Mol Biol 215:403–410

The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408:796–815

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

Ciceri P, Locatelli F, Genga A, Viotti A, Schmidt RJ (1999) The activity of the maize Opaque2 transcriptional activator is regulated diurnally. Plant Physiol 121:1321–1327

Cord Neto G, Yunes JA, Vettore AL, da Silva MJ, Arruda P, Leite A (1995) The involvement of Opaque2 in β-prolamine gene regulation in maize and Coix suggests a more general role for this transcriptional activator. Plant Mol Biol 27:1015–1029

Damerval C, le Guilloux M (1998) Characterization of novel proteins affected by the *o2* mutation and expressed during maize endosperm development. Mol Gen Genet 257:354–361

Dayhoff MO, Schwartz RM, Orcutt BC (1978) In: Dayhoff MO (ed) Atlas of protein sequence and structure, Vol 5, Suppl 3. National Biochemical Research Foundation, Silver Spring, MD, pp 345–352

Felsenstein J (1993) PHYLIP (phylogeny inference package) version 3.5c. Distributed by the author. Department of Genetics, University of Washington, Seattle, WA

Fitch WM (2000) Homology. A personal view on some of the problems. Trends Genet 16:227–231

Gallusci P, Varrot S, Matsuoko M, Maddaloni M, Thompson RD (1996) Regulation of cytosolic pyruvate, orthophosphate dikinase expression in developing maize endosperm. Plant Mol Biol 31:45–55

Giraudat J, Hauge BM, Valon C, Smalle J, Parcy F, Goodman HM (1992) Isolation of the Arabidopsis ABI3 gene by positional cloning. Plant Cell 4:1251–1261

Goff AS, Ricke D, Lan T-H, et al. (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonicum*). Science 296:92–100

Graur D, Li W-H (2000) Fundamentals of molecular evolution. Sinauer Associates, Sunderland, MA

Green P (1994) *phrap* (http://www.genome.washington.edu/)

Gu X (1999) Statistical methods for testing functional divergence after gene duplication. Mol Biol Evol 16:1664–1674

Hughes AL (1994) The evolution of functionally novel proteins after gene duplication. Proc R Soc Lond B 256:119–124

Hurst H (1995) Transcription factor 1: bZIP proteins. Protein Profile 2:105–168

Jakoby M, Weisshaar B, ge-laser W, Carbajosa JV, Tiedemann J, Kroj T, Parcy F (The bZIP Research Group) (2002) bZIP transcription factors in Arabidopsis. Trends Plant Sci 7:106–111

Kellogg EA (2001) Evolutionary history of the grasses. Plant Physiol 125:1198–1205

Kemper EL, Cord Neto G, Papes F, Martinez Moraes KC, Leite A, Arruda P (1999) The role of Opaque2 in the control of lysine-degrading activities in developing endosperm. Plant Cell 11:1981–1993

Kircher S, Wellmer F, Nick P, Rügner A, Schäfer E, Harter K (1999) Nuclear import of the parsley bZIP transcription factor CPRF2 is regulated by phytochrome photoreceptors. J Cell Biol 144:201–211

Kumar S, Tamura K, Jakobsen IB, Nei M (2000) Molecular evolutionary genetic analysis (MEGA) v 2.1. http://www.mega-software.net/

Li WH (1993) Unbiased estimation of the rates of synonymous and nonsynonymous substitutions. J Mol Evol 36:96–99

Lohmer S, Maddaloni M, Motto M, Di Fonzo N, Hartings H, Salamini F, Thompson RD (1991) The maize regulatory locus Opaque-2 encodes a DNA-binding protein which activates the transcription of the b-32 gene. EMBO J 10:617–624

Lupas A (1996) Prediction and analysis of coiled-coil structures. Methods Enzymol 266:513–525

Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. Science 290:1151–1155

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

Nadeau JH, Sankoff FD (1997) Comparable rates of gene loss and functional divergence after genome duplication early in vertebrate evolution. Genetics 147:1259–1266

Ohno S (1970) Evolution by gene duplication. Spring Verlag, Berlin–Heidelberg–New York

Ohta T (2000) Evolution of gene families. Gene 259:45–52

Oñate L, Vicente-Carbajosa J, Lara P, Díaz I, Carbonero P (1999) Barley BLZ2, a seed-specific bZIP protein that interacts with BLZ1 in vivo and activates transcription from the GCN4-like motif of B-hordein promoters in barley endosperm. J Biol Chem 274:9175–9182

Onodera Y, Suzuki A, Wu C-Y, Washida H, Takaiwa F (2001) A rice functional transcription activator, RISBZ1, responsible for endosperm-specific expression of storage protein genes through GCN4 motif. J Biol Chem 276:14139–14152

Pamilo P, Bianvhi NO (1993) Evolution of the Zfx and Zfy genes: Rates and interdependence between the genes. Mol Biol Evol 19:271–281

Purugganan MD, Wessler SR (1994) Molecular evolution of the plant R regulatory gene family. Genetics 138:849–854

Pysh LD, Schmidt RJ (1996) Characterization of the maize OHP1 gene: Evidence of gene copy variability among inbreds. Gene 177:203–208

Pysh LO, Aukerman MJ, Schmidt RJ (1993) OHP1: A maize basic domain/leucine zipper protein that interacts with Opaque2. Plant Cell 5:227–236

Riechmann JL, Heard J, Martin J, Reuber F, Jiang CZ, Keddie J, Adam L, Pineda O, Ratcliffe OJ, Samaha RR, Creelman R, Pilgrim M, Broun P, Zhang JZ, Ghanderhari D, Sherman BK, Yu GL (2000) Arabidopsis transcription factors: Genome-wide comparative analysis among eukaryotes. Science 290:2105–2110

Rost B (1996) PHD: Predicting one-dimensional protein structure by profile based neural networks. Methods Enzymol 266:525–539

Sambrook J, Fritsch EF, Maniatis T (1989) Molecular cloning: A laboratory manual. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY

Schmidt RJ (1993) Opaque-2 and zein genes expression. In: Verna DPS (ed) Control of plant gene expression. CRC Press, Boca Raton, FI, pp 337–355

Schmidt RJ, Ketudat M, Aukerman MJ, Hoschek G (1992) Opaque-2 is a transcriptional activator that rePCOGnizes a specific target site in 22-kD zein genes. Plant Cell 4:689–700

Schmitz D, Lohmer S, Salamini F, Thompson RD (1997) The activation domain of the maize transcription factor Opaque-2 resides in a single acidic region. Nucleic Acids Res 25:756–763

Shewry PR, Napier JA, Tatham AS (1995) Seed storage proteins: Structures and biosynthesis. Plant Cell 7:945–956

Tajima F (1993) Simple methods for testing molecular clock hypothesis. Genetics 135:599–607

Tatusov RL, Koonin EV, Lipman DJ (1997) A genomic perspective on protein families. Science 278:631–637

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, 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

Thornton JW, DeSalle R (2000) Gene family evolution and homology: genomics meets phylogenetics. Annu Rev Genomics Hum Genet 1:41–73

Varagona MJ, Raikhel NV (1994) The basic domain in the bZIP regulatory protein Opaque2 serves two independent functions: DNA binding and nuclear localization. Plant J 5:207–214

Vettore AL, Yunes JA, Cord Neto G, da Silva MJ, Arruda P, Leite A (1998) The molecular and functional characterization of an Opaque2 homologue gene from *Coix* and a new classification of plant bZIP proteins. Plant Mol Biol 36:249–263

Vicente-Carbajosa J, Moose SP, Parsons RL, Schmidt RJ (1997) A maize zinc-finger protein binds the prolamine box in zein gene promoters and interacts with the basic leucine zipper transcriptional activator Opaque2. Proc Natl Acad Sci USA 94:7685–7690

Vicente-Carbajosa J, Oñate L, Lara P, Diaz I, Carbonero P (1998) Barley BLZ1: A bZIP transcriptional activator that interacts with endosperm-specific gene promoters. Plant J 13:629–640

Vincentz M, Schlögl P, Corrêa LG, Kühne F, Leite A (2001) Phylogenetic relationships between Arabidopsis and sugarcane bZIP transcriptional regulatory factors. Gen Mol Genet 24:55–60

Vision TJ, Brown DG, Tanksley SD (2001) The origin of genomic duplications in Arabidopsis. Science 290:2114–2117

Walsh JB (1995) How often do duplicated genes evolve new functions. Genetics 139:421–428

Wendel JF (2000) Genome evolution in polyploids. Plant Mol Biol 42:225–249

Wingender E, Chen X, Hehl R, Karas H, Liebich I, Matys V, Meinhardt T, Prüß M, Reuter I, Schacherer F (2000) TRANSFAC: An integrated system for gene expression regulation. Nucleic Acids Res 28:316–319

Yang YW, Lai KN, Tai PY, Li WH (1999) Rates of nucleotide substitution in angiosperm mitochondrial DNA sequences and dates of divergence between Brassica and other angiosperm lineages. J Mol Evol 48:597–604

Yu J, Hu S, Wang J, et al. (2002) A draft sequence of the rice genome (*Oyza sataiva* L. ssp. *Indica*). Science 296:79–91

Yunes JÁ, Vettore AL, da Silva MJ, Leite A, Arruda P (1998) Cooperative DNA binding and sequence discrimination by the Opaque2 bZIP fector. Plant Cell 10:1941–1955

Zhang J, Rosenberg HF, Nei M (1998) Positive darwinian selection after gene duplication in primate ribonuclease gene. Proc Natl Acad Sci USA 95:3708–3713

Zhu X, Tang G, Granier F, Bouchez D, Galili G (2001) A T-DNA insertion knockout of the bifunctional lysine-ketoglutarate reductase/saccharopine dehydrogenase gene elevates lysine levels in Arabidopsis seeds. Plant Physiol 126:1539–1545