Phylogeny, Function, and Evolution of the Cupins, a Structurally Conserved, Functionally Diverse Superfamily of Proteins

Sawsan Khuri, Freek T. Bakker,[1] *and Jim M. Dunwell*

School of Plant Sciences, University of Reading, Reading, England

The cupin superfamily is a group of functionally diverse proteins that are found in all three kingdoms of life, Archaea, Eubacteria, and Eukaryota. These proteins have a characteristic signature domain comprising two histidine-containing motifs separated by an intermotif region of variable length. This domain consists of six beta strands within a conserved beta barrel structure. Most cupins, such as microbial phosphomannose isomerases (PMIs), AraC-type transcriptional regulators, and cereal oxalate oxidases (OXOs), contain only a single domain, whereas others, such as seed storage proteins and oxalate decarboxylases (OXDCs), are bi-cupins with two pairs of motifs. Although some cupins have known functions and have been characterized at the biochemical level, the majority are known only from gene cloning or sequencing projects. In this study, phylogenetic analyses were conducted on the conserved domain to investigate the evolution and structure/function relationships of cupins, with an emphasis on single-domain plant germin-like proteins (GLPs). An unrooted phylogeny of cupins from a wide spectrum of evolutionary lineages identified three main clusters, microbial PMIs, OXDCs, and plant GLPs. The sister group to the plant GLPs in the global analysis was then used to root a phylogeny of all available plant GLPs. The resulting phylogeny contained three main clades, classifying the GLPs into distinct subfamilies. It is suggested that these subfamilies correlate with functional categories, one of which contains the bifunctional barley germin that has both OXO and superoxide dismutase (SOD) activity. It is proposed that GLPs function primarily as SODs, enzymes that protect plants from the effects of oxidative stress. Closer inspection of the DNA sequence encoding the intermotif region in plant GLPs showed global conservation of thymine in the second codon position, a character associated with hydrophobic residues. Since many of these proteins are multimeric and enzymatically inactive in their monomeric state, this conservation of hydrophobicity is thought to be associated with the need to maintain the various monomer-monomer interactions. The type of structure-based predictive analysis presented in this paper is an important approach for understanding gene function and evolution in an era when genomes from a wide range of organisms are being sequenced at a rapid rate.

Introduction

The recently designated cupin superfamily of proteins (Dunwell 1998*a*) offers a model with which to investigate evolutionary relationships among proteins of similar structures but very different functions (Aravind and Koonin 1999*a*). The conserved domain, comprising a six-stranded beta barrel structure (Gane, Dunwell, and Warwicker 1998; Woo et al. 2000), was given the name cupin (from the Latin word *cupa,* meaning ''small barrel''). Extensive sequence analysis has identified representatives of this type of protein in all prokaryotic and eukaryotic lineages examined, with the possible exception of the spirochaetes (Dunwell, Khuri, and Gane 2000).

The cupin superfamily was discovered when amino acid sequence analysis, originally based on the nonapeptide ''germin-box'' (Lane et al. 1991), found a high level of similarity between fungal spherulins, produced upon spore formation in the slime mold *Physarum polycephalum,* and wheat germin, a thermostable, glycosylated protein produced during germination and shown to have oxalate oxidase (OXO) (EC 1.2.3.4) activity (Lane et al. 1993; Lane 2000). This small family of germin-related proteins was then found to share a small number of globally conserved residues with 7S (vicilin) and 11S (legumin) seed storage proteins (Bäumlein et al. 1995). Subsequent analysis (Dunwell and Gane 1998), using an extended version of the germin-box, identified the conserved domain in a large number of microbial proteins. These included type II phosphomannose isomerases (PMIs) (EC 5.3.1.8), found in a large number of prokaryotes, AraC-type (Gallegos et al. 1997) and other (Aravind and Koonin 1999*b*) transcription regulators, gentisate 1,2-dioxygenases (GDOs) (EC 1.13.11.4), and oxalate decarboxylases (OXDCs) (EC 4.1.1.2).

Although some functional information is available, the majority of bacterial and archaeal cupins, as well as an increasing number from plants (generally known as germin-like proteins [GLPs]) and animals, have been identified from genome sequencing projects and are of unknown function. Together with the experimental confirmation of the bifunctional nature of barley germin as an enzyme with both OXO and manganese superoxide dismutase (Mn-SOD) (EC 1.15.1.1) activity (Woo et al. 2000), as well as the recent evidence that a GLP from moss (Yamahara et al. 1999) and one from tobacco (Carter and Thornburg 2000) are Mn-SODs, there is circumstantial evidence relating to the possible functions of many other germins and GLPs. Some are expressed at critical developmental stages such as embryogenesis (Domon et al. 1995; Neutelings 1998) or floral induction (Heintzen et al. 1994; Staiger, Apel, and Trepp 1999), and many are induced by a range of stresses, either biotic, such as infection with powdery mildew (Thordahl-

[1] Present address: Wageningen University, Plant Taxonomy Group, Wageningen, the Netherlands.

Key words: phylogeny, cupin, germin, oxalate oxidase, phosphomannose isomerase, protein structure.

Address for correspondence and reprints: Jim M. Dunwell, School of Plant Sciences, University of Reading, Reading RG6 6AS, United Kingdom. E-mail: j.m.dunwell@reading.ac.uk.

Mol. Biol. Evol. 18(4):593–605. 2001

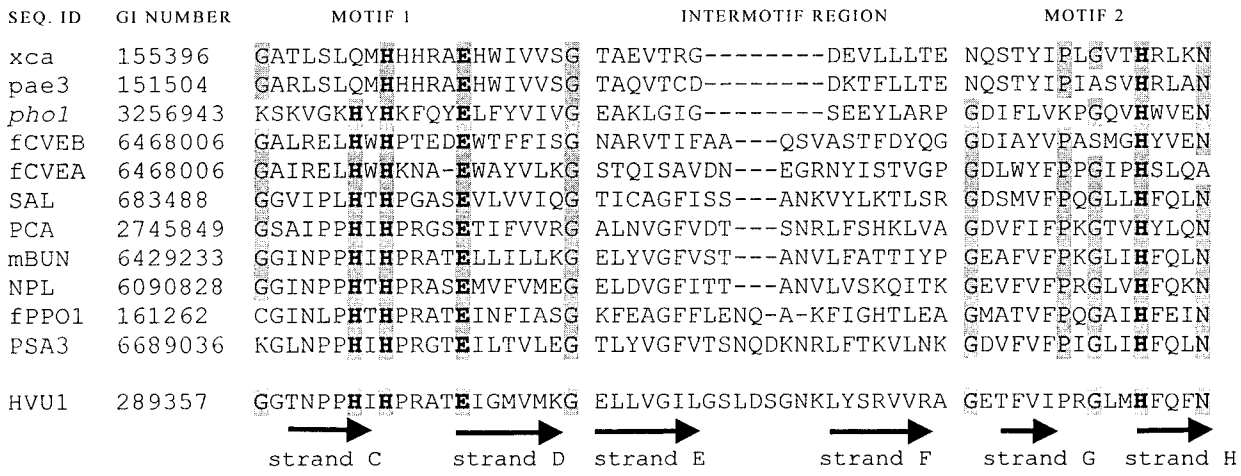

FIG. 1.—Alignment of the conserved domain in a sample of typical cupin sequences from archaea (*pho1*), bacteria (xca, pae3), fungi (fCVEA and B, fPPO1), moss (mBUN), gymnosperms (PCA), and higher plants (SAL, NPL, PSA3, and HVU1). The consensus cupin signature residues are highlighted, the metal-binding active site ligands of three histidines and a glutamate are in bold, and the arrows denote the six β-strands.

Christensen et al. 1997; Schweizer, Christoffel, and Dudler 1999), or abiotic, such as exposure to salt (Hurkman and Tanaka 1996), aluminum (Hamel, Breton, and Houde 1998), or high temperatures (Vallelian-Bindschedler et al. 1998).

Cupins have two histidine-containing motifs (fig. 1), corresponding to the C/D and G/H strands of OXO (Woo et al. 2000), separated by an intermotif region (IMR) which includes strands E and F and the intervening loop. The IMR of the single domain cupins varies in length from 15 amino acids in most archaeal and bacterial cupins to 26 residues in some of the cereal proteins (Dunwell, Khuri, and Gane 2000). It is presumed that this variation in intermotif length, due principally to insertions in the interstrand loop, delayed the identification of the conserved residues in this diverse superfamily of proteins. It is now known that the two conserved histidines and a glutamate in the first motif, together with a third conserved histidine in the second motif (fig. 1), act as ligands for the active-site metal, which has been shown to be a single manganese atom for barley germin (Requena and Bornemann 1999; Woo et al. 2000), and for the GLPs from a moss (*Barbula unguiculata*; Yamahara et al. 1999) and from tobacco (Carter and Thornburg 2000).

One diagnostic feature that can be used to discriminate between the various classes of cupin is whether the conserved domain occurs singly, in proteins such as transcription regulators and PMIs (Dunwell 1998*a*), or in a duplicated form within proteins such as the OXDCs (Dunwell and Gane 1998) and seed storage proteins (Bäumlein et al. 1995); these two-domain proteins have been termed bi-cupins (Dunwell 1998*a*). The evolutionary origin of cupins, and, indeed, that of bi-cupins, has yet to be resolved; these latter could have been the result of one or more gene duplication events leading to separate evolution of the different classes (Shutov, Blattner, and Baumlein 1999). This paper presents evidence on the evolutionary relationships within the cupin superfamily, with a focus on single-domain plant GLPs, the conservation of structure, and the significance of the active-site ligands in the possible function of these proteins.

Materials and Methods

In view of the extensive variation in overall length of the various cupin proteins, and as a consequence of preliminary analyses, the sequences used in the present study were restricted to the conserved cupin domain that comprises the two conserved motifs [$\mathbf{G}(X)_5\mathbf{H}X\mathbf{H}(X)_{3,4}\mathbf{E}(X)_6\mathbf{G}$] and $\mathbf{G}(X)_5\mathbf{P}X\mathbf{G}(X)_2\mathbf{H}(X)_3\mathbf{N}$], together with the intermotif region (IMR) (fig. 1).

The various sequences were identified using BLAST searches with a variety of cupin sequences from different species. The nonredundant, nonmouse, and nonhuman EST and microbial GenBank databases at the National Institute for Biotechnology Information (NIH, Bethesda, Md.) were mainly used via http://www.ncbi.nlm.nih.gov/BLAST; other sequences were identified by using the Institute of Genome Research TIGR database (Rockville, Md.) at http://www.tigr.org/cgi-bin/BlastSearch/blast.cgi and the Sanger Centre database (Cambridge, England) at http://sanger.ac.uk. To avoid errors and misleading results, all microbial sequences used in these analyses were from completed genomes, except for the *Streptococcus mutans* bi-cupin putative OXDC. Similarly, identical cupin sequences (e.g., from *Arabidopsis thaliana*) were deleted, and only a small number of plant ESTs were used in the phylogenetic studies. General alignments were performed using CLUSTAL W at the BCM Search Launcher at http://dot.imgenn.bcm.tmc.edu:9331/ and manually adjusted to produce a largely unambiguous alignment. Subsequent phylogenetic analysis was performed using PAUP*, version 4.0b4a (Swofford 1999). A global alignment was constructed with a total of 120 protein sequences comprising 25 plant GLPs, 5 fungal sequences, and 90 microbial sequences, of which 13 were from archaeal species. A second separate alignment was con-

structed, containing 73 GLPs from plants and fungi. Results of the unrooted global analysis indicated that ePBA was an appropriate root for the GLP analysis (see below).

As described above, phylogenetic analyses were performed for both alignments with the IMR region included; excluding this region from the analyses resulted in loss of resolution in the topologies found. Parsimony searches in both analyses involved heuristic searches with the following settings: gaps were treated as ''missing''; branches were collapsed if maximum branch length was zero; topological constraints were not enforced, and STEEPEST DESCENT, ACCTRAN, and MULPARS were in effect. In order to measure clade support, jackknife analysis was carried out using PAUP with settings emulating Parsimony Jackknifer (Farris 1969), i.e., percentage of characters deleted in each replicate = 37, ''fast'' stepwise addition, ''Jac'' resampling, and only groups with jac frequency >50% kept. Initial (equally weighted) searches were constrained by jackknife trees thus obtained so as to avoid searching through irrelevant parts of tree space. Given the expectation that the sequences in our data sets could be highly divergent, the ''Protpars'' stepmatrix as implemented in MacClade (version 3.0.8a; Maddison and Maddison 1992) was used in subsequent searches in order to take into account differences in probabilities of substitutions among amino acids (i.e., stepmatrix values are based on the minimum number of changes between amino acids, calculated from the genetic code).

The 120-protein sequence alignment contained 70 positions, of which 25 were part of the IMR. In total, the alignment contained 65 variable positions, of which 62 were phylogenetically informative. A heuristic parsimony search was performed with 100 random sequence additions (''hold'' = 3) without branch swapping, with ''MULPARS off,'' and saving trees even when they were not optimal overall. The resulting trees were then used as starting trees in a search with ''MULPARS on'' and TBR branch swapping, which yielded >20,300 most-parsimonious trees (MPTs) of length 2,039 (consistency index [CI] = 0.35, retention index [RI] = 0.57). The search was repeated with the Protpars stepmatrix implemented in order to reflect generally observed substitution patterns in amino acid sequences. This yielded 6 MPTs of length 2,688 (CI = 0.26, RI = 0.59), of which the strict consensus, with corresponding jackknife support values, was taken as the final result. This stepmatrix-weighted strict consensus tree was largely congruent with and better resolved than the unweighted strict consensus (not shown).

The sequence alignment of 73 plant GLPs contained 64 positions, of which 53 were phylogenetically informative. A heuristic parsimony search with 50 replicates of random sequence addition (holding three trees per step), with TBR branch swapping, and keeping only trees compatible with a jackknife tree calculated from the same data set yielded 3,660 MPTs of length 631 (CI = 0.49, RI = 0.75). Repeating the analysis with the Protpars stepmatrix implemented and without topological constraints yielded 2,928 MPTs of weighted length 761 (CI = 0.42, RI = 0.76), of which the strict consensus tree, rooted with the ''ePBA'' clade, is presented below. This rooting was based on the global topology (see above). As with the global analysis, the effect of implementing the Protmars stepmatrix was a further increase in resolution in the plant-weighted strict consensus topology compared with that from the unweighted analysis (not shown).

Each sequence used in this study is denoted by a three-letter code made up of the first letter of the generic name and the first two letters of the epithet, with prokaryotic abbreviations in lowercase and eukaryotic ones in uppercase; e.g., ''pae'' denotes *Pseudomonas aeruginosa*, and ''HVU'' denotes *Hordeum vulgare*. Tables 1 and 2 contain a complete listing of the sequences used.

Results and Discussion

Global Analysis

Phylogenetic Analysis

The conserved cupin domain section of the protein sequences used in these analyses contained enough phylogenetic signal to allow inference of evolutionary and functional relationships among them. The phylogeny presented in figure 2 shows that the plant GLP sequences form a separate clade and that cupins from proteins of similar, or the same, putative functions generally cluster together. While it is recognized that branching order within main clades could change upon further taxon sampling, it is expected that the main clades themselves will remain identifiable.

Most of the hypothesized functions of sequences in the gene databases are based only on sequence similarity and not on biochemical evidence (see table 1). Thus, any functional relationships that are inferred from this phylogeny are made in accordance with these assumptions. The functional groups that resulted from this analysis (fig. 2) were the bacterial and archaeal PMI and PMI-like sequences, the two bacterial GDOs, and the fungal OXDCs which formed a group together with prokaryotic sequences of putative OXDC function. The AraC-type and other transcription regulators, however, were dispersed.

The other sequences used were of unknown function. Several changes in IMR length have occurred across the phylogeny (fig. 2), from the basic, ancestral value of 15 residues found in the PMIs and transcriptional regulators to a maximum of 26 in a small number of cereal GLPs (see below).

The PMI Cluster

Within the monophyletic PMI clade (fig. 2), there are two key observations to note. First, the *A. aeolicus* (gi|2983213, aae) sequence is on a relatively long branch derived from within the PMI group, a pattern consistent with hypotheses about the ancient origin of this bacterium (Deckert et al. 1998). This cupin has only two conserved histidines; the glutamate and third histidine that complete the putative active site are substituted by valine and tyrosine, respectively. There is only a single nucleotide difference between glutamate (GAA/G) and

Table 1
List of Microbial Cupin Sequences Used with Species Names, GenBank Identifiers (gi), Gene Names, and Functions Where Known

Organism	gi No.	Gene	Function	Abbreviation
Acineobacter calcoaceticus	559389	*epsM*	~GMP + PMI	aca
Acetobacter xylinus	2569942	*aceF*	~PMI	axy
Anacystis nidulans	1405429	*orf150*	?	ani
Aquifex aeolicus	2983213	*xanB*	~GMP + PMI	aae
Archaeoglobus fulgidus	2649385	AF1208	HP	*afu3*
	2649495	AF1097	~PMI	*afu2*
	2689405	AF0326	~PMI	*afu1*
Azotobacter vinelandii	2231995	*aldh*	ALDH	avi
Bacillus subtilis	2632741	*ydbB*	?	bsu2
	1881251	*ydbB*	?	bsu3
	1881323	*ydeC*	~TR	bsu1
	563940	*yxaG*	?	bsu5
	563940	*yxaG*	?	bsu4
	2635837	*yvrK*	~OXDC[b,c]	bsu6
	2619026	*yoaN*	?~OXDC[b]	bsu7
Erwinia chrysanthemi	48984	*kdgF*	?	ech
Escherichia coli	40936	*araC*	TR	eco2
	1155018	*manC*	GMP + PMI	eco6
	415623	*manC*	~GMP + PMI	eco9
	3142209	*manC*	~GMP + PMI	eco14
	305009	*rhaS*	TR	eco5
	305010	*rhaR*	TR	eco4
	3142217	*manC*	~GMP + PMI	eco11
	3142222	*manC*	~GMP + PMI	eco12
	1742129	*ycjC*	Immunity repressor	eco1
	598465	*rfbM*	GMP + PMI	eco10
	4867924	*manC*	~GMP + PMI	eco7
	441136	*rfbM1*	GMP ± PMI	eco13
	147164	*pmi*	PMI	eco8
	1742832	*chiB*	TR (*celD*)	eco3
Helicobacter pylori	2313118	HP0043	~PMI	hpy
Klebsiella pneumoniae	747676	—	?	kpn1
	3142227	*manC*	~GMP + PMI	kpn2
Listeria monocytogenes	2745844	*lapB*	~TR	lmo
Methylobacterium extorquens	2394390	—	~PMI	mex
Methanococcus jannaschii	1499583	MJ0764	HP	*mja2*
	1592216	MJ1618	?	*mja1*
Methanobacterium thermoautotrophicum	262140	MTH352	?	*mth4*
	2621742	MTH659	?	*mh2*
	2621786	MTH700	?	*mth1*
	2622642	MTH1522	?	*mth3*
Mycobacterium tuberculosis	2104294	Rv2618	?	mtu4
	2104394	Rv3471c	?	mtu3
	1781124	Rv3833	~TR	mtu2
	2213518	Rv0181c	?	mtu1
Pseudomonas sp. U2 plasmid	3406827	*nagI*	GDO	pae2
Pseudomonas aeruginosa	3249549	*wbpW*	GMP + PMI	pae4
	3510759	*orf488*	~GMP + PMI	pae5
	151504	—	GMP + PMI	pae3
	406101	—	~Heat shock regulator	pae1
Pyrococcus abyssi	5458652	—	~GMP + PMI	*pab*
Pyrococcus horikoshii	3256432	PH0047	HP	*pho2*
	3256943	PH0537	HP	*pho1*
	3257338	PH0925	~GMP + PMI	*pho3*
Rhizobium sp. NGR234 plasmid	2182298	*noeJ*	~GMP + PMI	rhi2
	2182567	*y4oW*	HP	rhi1
Rhodospirillum rubrum pKY1 plasmid	216730	*pssM*	GMP + PMI	rru
Salmonella enterica	47902	*manC*	GMP + PMI	sen3
	47655	*manC*	GMP + PMI	sen1
	47013	*manC*	~GMP + PMI	sen4
	154204	*manC*	~GMP + PMI	sen2
Salmonella typhimurium	47906	*rhaC2*	TR	sty
Shigella sonnei	3930205	*manC*	GMP + PMI	sso
Sphingomonas sp. *RW5*	3550667	*gtdA*	GDO	sph
Streptococcus mutans Contig241[a]	Positions 5819–5652		? ~OXDC[b]	smu
Streptomyces coelicolor	4158181	SC9B5.02	HP	sco4
	3449255	SC6G4.21	?	sco1
	5139587	SC6G9.14	?	sco3
	3127846	SC1A6.14	~TR	sco2

Table 1
Continued

Organism	gi No.	Gene	Function	Abbreviation
Streptomyces cyaneus	153225	—	Polyketide synthase	scy
Streptomyces glaucescens	153495	*tcmJ*	?	sgl
Strepomyces halstedii	153322	*ORFB*	?	sha
Streptomyces purpurascens	581719	*ORF 1*	?	spu
Synechocystis sp. PCC6803	1001180	*rfbM*	~PMI	syn3
	1652906	sll1163	HP	syn2
	1653078	slr2101	HP	syn1
	1652486	*manA*	~PMI	syn4
	1652630	sll1358	?~OXDC[b]	syn5
Thermotoga maritima	4981179	TM0656	HP	tma1
	4981537	TM1010	HP	tma2
	4981845	TM1287	HP	tma3
Plasmid R751 transposon Tn4321	1402862	*orf1*	~Polyketide cyclase	tn4x
Vibrio cholerae	1230580	*orf*	GMP + PMI	vch1
	48383	*rfbA*	~GMP + PMI	vch2
Xanthomonas campestris	155396	*xanB*	GMP + PMI	xca
Yersinia enterocolitica	1197654	*rfbM*	GMP + PMI	yen

NOTE.—~ = putative/similar to; HP = hypothetical protein; TR = transcription regulator; GMP = GDP-mannose pyrophosphorylase; PMI = phosphomannose isomerase; OXDC = oxalase decarboxylase; ALDH = aldehyde dehydrogenase; GDO = gentisate 1,2 dioxygenase.

[a] Unfinished genome.

[b] Oxalate decarboxylase, proposed in Dunwell, Khuri, and Gane (2000).

[c] Verified in Tanner and Bornemann (2000).

valine (GTN) codons and between those encoding histidine (CAC/T) and tyrosine (TAC/T).

Second, the position of the cyanobacterial *Anacystis nidulans* (gi|1405429, ani) sequence as sister to all the other PMIs is intriguing. The likely function of this sequence as a PMI was verified by its similarity (*e* value 0.001) to a *P. aeruginosa* PMI (gi|3510759, pae5), although its most similar neighbors were two archaeal sequences from *Pyrococcus horikoshii* (gi|3257338, *pho3*, 4e−5) and *Pyrococcus abyssi* (gi|5458652, *pab*, 8e−5). In fact, all the archaeal PMI-like sequences group together after ''ani,'' forming a sister clade to the rest of the PMIs. This placement of the archaeal PMI-like sequences between a gram-negative bacterium and the gram-positive clades concurs with theories on the chimeric origin of the archaeal genome (Koonin et al. 1997).

Other Functional Groupings

The OXDC bi-cupins that were used for this analysis formed one clearly identified monophyletic clade. They included two fungal proteins: fCVE (gi|6468006) from *Collybia velutipes* (now known as *Flammulina velutipes*) (Kesarwani et al. 2000), and fAPH from *Aspergillus phoenices* (sequence from Scelonge and Bidney [1998], as amended by Dunwell, Khuri, and Gane [2000]), along with putative OXDC sequences from *Bacillus subtilis* (gi|2635837, bsu6; gi|2619026, bsu7), Synechocystis (gi|1652630, syn5) (Dunwell 1998*b*), and *S. mutans* (Contig205, DNA positions 6978–7153 and 7518–7688, smu). These analyses were conducted before recent experimental evidence confirmed bsu6 to be an OXDC (Tanner and Bornemann 2000). This finding, coupled with the high level of sequence similarity between these proteins, lends further support to the prediction that bsu7, syn5, and smu are OXDCs.

Although the two cupin domains (A and B) of each OXDC were treated separately in the analysis, all OXDC sequences formed one clade within which the two domains clustered as sister groups (fig. 2). The most likely interpretation of this is that a duplication event occurred once in the lineage leading to the OXDC clade, followed by divergence of the sequences. However, further analysis is needed before a conclusive statement is made regarding the evolution of the bi-cupins.

The only other group of proteins with known function that did not group consistently together in our analyses were the transcription regulators (TRs) (fig. 2). Most of those used in this study belong to the bacterial AraC/XylS family delineated by Gallegos et al. (1997) on the basis of the high sequence similarity of their helix-turn-helix DNA-binding domains; the cupin motifs occur within the N-terminal effector-binding domains of these proteins. This arrangement contrasts with that in the other TRs from thermophilic archaea and bacteria (gi|2621786, *mth*1; gi|2621742, *mth2*; gi|4981179, tmar1), in which the cupin domain is located at the C-terminal end of the protein (Aravind and Koonin 1999*b*). Whereas these three sequences clustered closer together, most of the AraC-type sequences were more dispersed. Two rhamnose-binding TRs, namely, *Escherichia coli* rhaS (gi|305009, eco5) and *Salmonella typhimurium* rhaC2 (gi|47906, sty), did form a group, and a third rhamnose-binding TR, rhaR (gi|305010, eco4) from *E. coli*, was nearby. Similarly, an *E. coli* chitobiose-binding TR (gi|1742832, eco3) (formerly considered to bind cellobiose), grouped with a *B. subtilis* TR (gi|2632741, bsu2). It is therefore possible that this dispersal might be a result of the range of effectors (sugars or other compounds) involved in binding to the cupin domain.

The alignment provided in figure 3 shows additional evidence of the similarity between the TRs and

Table 2
List of Eukaryotic Cupin Sequences Used with Species Names, GenBank Identifiers (gi), and Protein Names

Species	gi No.	Protein	Abbreviation
Arabidopsis thaliana	1755152	GLP4	ATH1
	6714406	Glp	ATH10
	6648167	Glp	ATH13
	6646782	Glp	ATH14
	6143869	Glp	ATH15
	5679328	GLP7	ATH16
	2244819	Glp	ATH128
	1755177	GLP5	ATH2
	1755184	GLP3a	ATH25
	6899900	GLP10	ATH3
	1755162	GLP2a	ATH33
	6721170	Glp	ATH4
	4098968	GLP9	ATH41
	2129594	glp type 2	ATH43
	1934730	GLP10	ATH44
	1755182	GLP7	ATH46
	1755166	GLP6	ATH47
	6721168	Glp	ATH5
	3047078	Glp	ATH51
	1592672	Germin1	ATH52
	6721167	Glp	ATH6
	6721166	Glp	ATH7
	6721165	Glp	ATH8
	6714408	Glp	ATH9
Aspergillus phoenices	NA	—	fAPH
Atriplex lentiformis	4996622	Glp	ALE
Barbula unguiculata	6429233	Partial[a]	mBUN
Brassica napus	914910	Glp	BNA
Ceratodon purpureus	6068694	EST	emCPU
Collybia velutipes	6468006	OXDC[b]	fCVE
Fragaria x ananassa	NA	Glp	FAN
Glycine max	4289984	EST	eGMA
Gossypium hirsutum	5047806	EST	eGHI
	6650732	glp1	GHI
Hordeum vulgare	289357	OXO[a,c] (1993)	HVU1
	1171937	OXO[a,c] (1994)	HVU2
	2815292	Glp	HVU3
	1070358	Glp	HVU4
Lycopersicon esculentum	2979494	?	LES
Medicago truncatula	6101843	EST	eMTR
Mesembryanthemum crystallinum	167258	Glp	MCR
Nicotiana plumbaginifolia	6090828	NEC1[a]	NPL
Oryza sativa	2655291	GER4	OSA1
	2655289	GER3	OSA11
	2655287	GER2	OSA12
	2655285	GER1	OSA13
	2801803	Glp16	OSA14
	5499730	RGLP1	OSA2
	5042461	Glp	OSA3
	5852087	Glp	OSA4
	599732	RGLP2	OSA5
	4239821	Osglp1	OSA6
	3293559	GER7	OSA7
Pharbitis nil	662292	Glp	PNI
Physarum polycephalum	161262	Spherulin 1A	fPPO1
	134860	Spherulin 1B	fPPO2
Pinus caribaea	2745849	PcGER1	PCA
Pinus radiata	2935521	PRGER1	PRA
	2739260	Partial	PSA1
	6689036	Ger2a	PSA3
	668903	Ger1	PSA4
Populus balsamifera	3857819	EST	ePBA
Prunus persica	4098517	ABP19[d]	PPE1
	1916809	ABP19[d]	PPE2
	1916807	ABP20[d]	PPE3
Sinapis alba	683488	Glpl	SAL
Solanum tuberosum	3171251	OXAOXA	STU
Triticum aestivum	170698	gf-2.8[c]	TAE1

Table 2
Continued

Species	gi No.	Protein	Abbreviation
	1772597	PSBGer1	TAE10
	121131	gf-3.8	TAE2
	5869975	glp2b	TAE3
	5869973	glp2a	TAES4
	1772601	pSBGer3	TAE8
	1772599	pSBGer2	TAE9
Zea mays.	4152154	EST	eZMA

NOTE.—NA = not available.
[a] Mn-superoxide dismutase.
[b] Oxalate decarboxylase.
[c] Oxalate oxidase.
[d] Auxin-binding protein.

highlights three points. First, the AraC-type group can be distinguished clearly from the thermophilic group in having only one of the conserved histidine residues in motif 1. Second, the archetypal arabinose-binding *E. coli* AraC (gi|40936, eco2), although similar to other cupins in terms of its beta-barrel structure (Soisson et al. 1997; Dunwell, Khuri, and Gane 2000), is not typical of its group in terms of sequence. Third, the two sequences from *B. subtilis* (gi|1881323, bsu1) and *Mycobacterium tuberculosis* (gi|2213518, mtu2) both have a glutamine residue in place of the glutamate in the putative binding site of these proteins. A more comprehensive alignment of 112 TRs and putative TR sequences showed that a total of 54 proteins with a glutamine at that position compared with 33 that had a glutamate residue (data not shown). This contrasts with the higher conservation frequencies of the three histidines, which were 94, 95, and 104, respectively.

As expected for the majority of the eukaryotic cupins used, the plant sequences formed a monophyletic group which included the fungal sequence used, *P. polycephalum* spherulin (gi|161261, fPPO1) (fig. 2). The best characterized of the plant GLPs are the germins from barley (gi|289356, HVU1) and wheat (gi|170697, TAE1). The detailed analysis of the plant and fungal GLPs is provided below.

Plant and Fungal Germin-like Proteins

Phylogenetic Analysis

The plant GLPs used in this study included examples from most main plant lineages, including cereals, legumes, gymnosperms, and two moss GLPs (*B. unguiculata*, gi|6429233, mBUN; *Ceratodon purpureus* EST gi|6068694, emCPU). In addition, two fungal spherulins from *P. polycephalum* (gi|161262 and gi|134860; fPPO1 and 2) were included. The resulting phylogeny contained three main groups (fig. 4), a result which differs from that recently proposed by Carter and Thornburg (1999, 2000), who identified five separate clades. However, Carter and Thornburg produced their phylogeny using the whole protein sequence, rather than only the cupin domain. Furthermore, although the total number of sequences used for the phylogenies was coincidentally the same, Carter and Thornburg used 30 sequences from *A. thaliana*, 10 from *Oryza sativa* and 33 GLPs from other plant species, including one moss, whereas this analysis utilized 24 from *A. thaliana*, 11 from *O. sativa*, and 36 from other plant species, including (see above) two moss species and two from a fungus (table 2). The phylogeny presented in figure 4, therefore, gives a more complete picture of the evolutionary, as well as the functional, relationships among GLPs.

The first group (subfamily 1 [s/f 1]) can be classified as the ''true germin'' (Carter and Thornburg 1999) subclade and consists of the wheat and barley germins (gi|170698, TAE1; gi|289357, HVU1) along with some other GLPs of those species. Evidence is now accumulating that the true germins may be bifunctional enzymes with both SOD and OXO activity (Woo et al. 2000).

Subfamily 2 (s/f 2) includes GLPs from a wide range of taxonomic groups, including other cereals, gymnosperms, and halophytic species, namely *Atriplex lentiformis* (gi|4996622, ALE) and *Mesembryanthemum crystallinum* (gi|167258, MCR). Furthermore, these cupins were isolated from diverse and seemingly unconnected sources. The solanaceous group, for example, consisted of GLPs from Mn-deficient roots of tomato (gi|2979494, LES), nectar of tobacco (gi|6090828, NPL), and cell suspension cultures of potato (gi|3171251, STU). The two gymnosperm GLPs were both isolated from somatic embryos (Neutelings 1998).

Despite this diversity of origin, nested within this subfamily is a group of proteins linked directly to stress tolerance: MCR was isolated from salt-stressed tissue, and OSA2 (gi|5499730) and OSA5 (gi|5499732) were labeled stress-responsive root GLPs. In addition, two centrally positioned proteins within s/f 2 are now known to have Mn-SOD activity, namely NPL (gi|6090828) from *Nicotiana plumbaginifolia* (Carter and Thornburg 2000) and mBUN (gi|6429233) from the moss *B. unguiculata* (Yamahara et al. 1999). It is therefore possible to predict that the GLPs in s/f 2 all have a role in tolerance to oxidative stress and are most likely to be Mn-SODs. This is in accordance with the predictions of Carter and Thornburg (2000), and it is expected that more experimental evidence will be forthcoming which will confirm these suggestions.

The third grouping, subfamily 3 (s/f 3), includes low-affinity auxin-binding proteins from peach

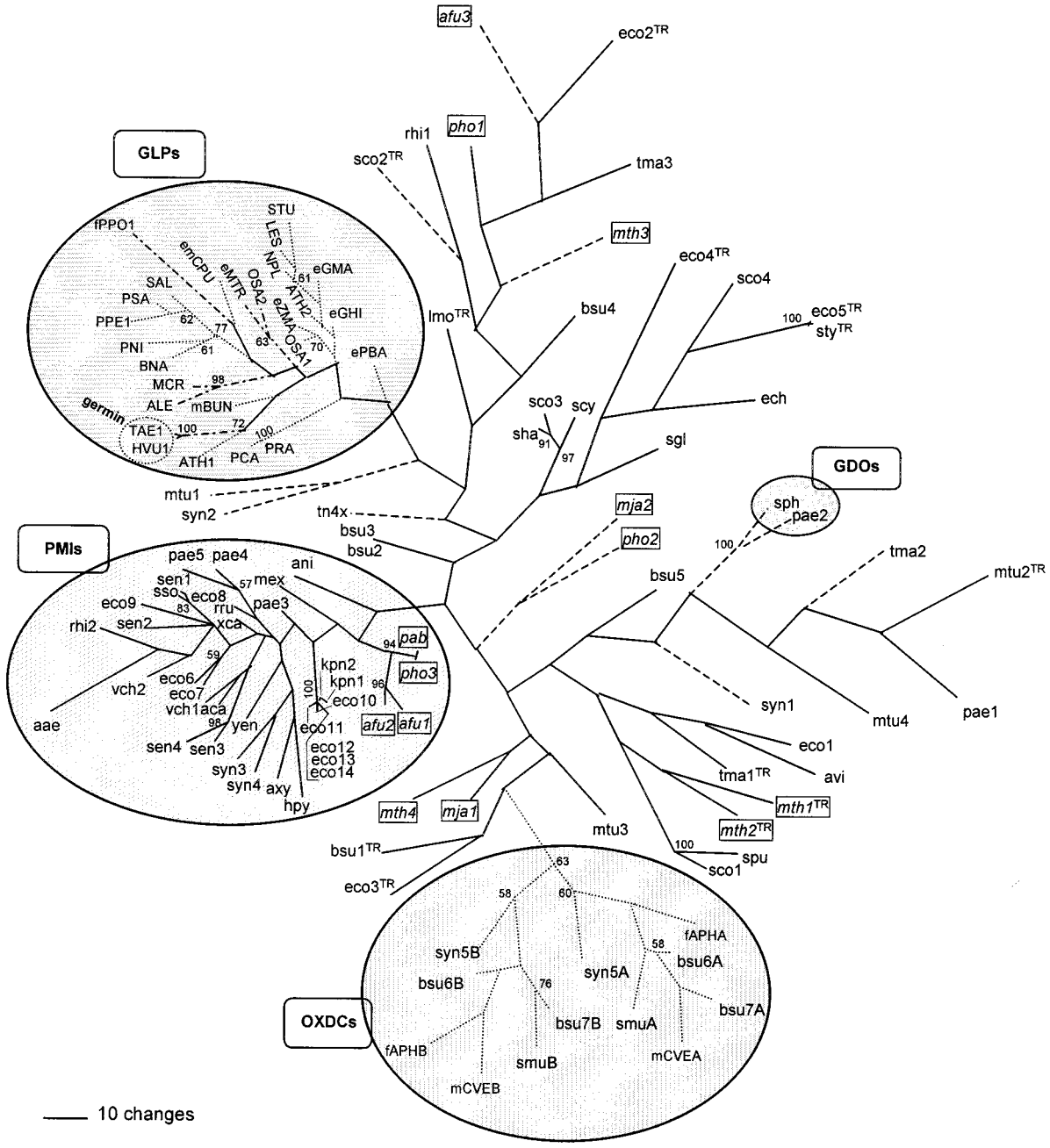

FIG. 2.—Global phylogeny (unrooted analysis): strict consensus of six most-parsimonious trees after Protpars stepmatrix character state weighting (see text). Numbers at nodes indicate jackknife values (10,000 replicates). GLP = germin-like protein; PMI = phosphomannose isomerase; OXDC = oxalate decarboxylase; GDO = gentisate 1,2 dioxygenase. Plant and fungal sequences are in uppercase letters; archaeal ones are italicized and boxed. Sequences have intermotif regions of 15 amino acids, with the following exceptions: dashed line = 16 (sco2 has 17); dotted line = 20; dot-and-dash line = 21 or more. TR = transcription regulator; prefix f = fungal; m = moss; e = EST; suffix A and B = first (N-terminal) and second (C-terminal) domains of OXDC, respectively. Full species names and accession numbers of the sequences used are presented in tables 1 and 2.

(gi|4098517, PPE1; gi|1916809, PPE2; gi|1916807, PPE3), as well as GLPs known to have a role linked to circadian rhythms and floral induction in *A. thaliana* (gi|1755184, ATH25; Staiger, Apel, and Trepp 1999), *Sinapis alba* (gi|683488, SAL; Heintzen et al. 1994), and *Pharbitis nil* (gi|662292, PNI; Ono et al. 1996). This subfamily may therefore consist mainly of regulatory proteins involved with auxin metabolism, either directly or indirectly.

The Metal-Binding Active Site

As seen above, with few exceptions, the enzymatic function(s) of GLPs remain largely unknown. It is clear

SEQ. ID	GI NUMBER	EFFECTOR	MOTIF 1	INTERMOTIF REGION	MOTIF 2
eco5	305009	rhamnose	QADFPE**H**H**H**DFH**E**IVIVEHG	TGIHVFNGQPYTITG	GTVCFVRDHDR**H**LYEH
sty	47906	rhamnose	QAAFPE**H**H**H**DFH**E**IVIVEHG	TGIHVFNGQPYTISG	GTVCFVRDHDR**H**LYEH
sco4	4158181	unknown	GTATTP**H**S**H**EDH**E**HFYVVRG	SGHAEVDGERTRIAA	GDALVVGAHQR**H**HFEN
eco4	305010	rhamnose	QDVFAE**H**T**H**DFC**E**LVIVWRG	NGLHVLNDRPYRITR	GDLFYIHADDK**H**SYAS
eco3	1742832	chitobiose	ESISGL**H**Q**H**DYY**E**FTLVLTG	RYFQEINGKRVLLER	GDFVFI**P**LGSH**H**QSFY
bsu1	1881323	unknown	NGYIPL**H**W**H**DEIQFVLILKG	IALFQINEEKIEVHE	GDGLFINSGYL**H**MAEE
lmo	2745844	unknown	EVLVYP**H**W**H**KEI**E**IIYALKG	SLNLGINDMPIQLKE	GEIQVINGGDV**H**YFLA
sco2	3127846	unknown	DTTWTE**H**S**H**PWH**E**LLWNAHG	ASTAVTGSQVWCVTP	TLGLWM**P**AGQL**H**SASA
mtu2	1781124	unknown	GARIER**H**R**H**PSHQIVYPSAG	AVSVTTHAGTWITPV	NRAIWI**P**AGCW**H**QHKF
eco2	40936	arabinose	FFIDRPLGMKGYILNLTIRG	QGVVKNQGREFVCRP	GDILLF**P**PGEI**H**HYGR
mth1	2621786	unknown	EEKPKTNS**H**PGQ**E**FNYVLEG	RIKFYIHDNEIILNE	GDSIFFDSSYE**H**AMEA
mth2	2621742	unknown	TDDFKLSS**H**EGE**E**FIYVLEG	EIEVIYGQDRYLLSE	GDSIYYDSVVP**H**HLHA
tma1	4981179	unknown	AQTEEESY**H**EGS**E**FGFVIQG	RIDLYLDGKRYRLKE	GDCFYYKADKK**H**YVKN

FIG. 3.—Alignment of a sample of transcription regulators, with their effector sugars, where known. Residues are highlighted as in figure 1. The upper group includes the bacterial AraC-type sequences with sequence sco4 included for comparison (see fig. 1). The lower group comprises the three sequences from thermophilic archaea and bacteria (Aravind and Koonin 1999*b*).

that most of them contain a single metal-binding active site with three conserved histidine ligands and one glutamate ligand (fig. 1). The position of the active site in the protein is protected within the beta-barrel structure (Woo et al. 2000), and it has recently been shown that the barley germin with OXO and SOD activity (Requena and Bornemann 1999; Woo et al. 2000), and the moss (Yamahara et al. 1999) and tobacco (Carter and Thornbug 2000) GLPs with SOD activity are manganese-containing enzymes. Both of these enzymes generate hydrogen peroxide, OXO by the oxidation of the oxalate ion, and SOD by the dismutation of superoxide radicals. It is likely that the types of catalysis are similar in all metal-binding GLPs or that there is a group, possibly s/f 3 as delineated above, that contains a different metal and therefore has slightly different properties.

Interestingly, a small number of GLPs do not contain all four of the highly conserved active-site residues. Instead of the third histidine (coded by CAY) residue in the second motif, PNI (gi|662292) has an aspartate (GAY), while ATH4 (gi|6721170), ATH5 (gi|6721168), and ATH7 (gi|6721166) have a glutamine (CAR). The lack of the third histidine infers that these proteins may not bind a manganese atom, or that they bind it very loosely and only under certain pH conditions. The conserved glutamate (GAR) in the first cupin motif can be substituted, again, by either aspartate (GAY), as in TAE8 (gi|1772601), or glutamine (CAR), as in mCPU (gi|6068694), ATH6 (gi|6721167), and ATH4 (gi|6721170). The absence of the negatively charged glutamate makes it unlikely that these proteins would bind manganese, and it is not known whether the equally negatively charged aspartate would provide the same function. In some *Glycine max* ESTs (gi|5759981, gi|6725711, and gi|6566187), this position is occupied by a positively charged lysine (AAR). Since all of these changes are due to a single nucleotide difference, they may simply be mutations that maintain the structure of these proteins but alter the binding characteristics of the active site and thus change its function.

The Intermotif Region

In the plant and fungal single-domain GLPs examined to date, the intermotif region varies in length from the usual minimum of 20 (exceptionally, 19 in some legume sequences) to 26 amino acids in some barley and wheat sequences (fig. 4). There is a high level of conservation within the IMR across all of the cupin subfamilies identified above, at both the protein and the DNA levels. An alignment of the DNA sequences (fig. 5) encoding the GLPs used in the global analysis allows the identification of certain patterns within the region encoding the IMR. First, two GLPs (PPE1 gi|4098517 and PSA1 gi|2739260) show evidence of two compensatory frameshifts. Second, where the IMR is longer than 20 amino acids, the position of the additional residues is consistently within the loop between structural strands E and F, to either side of a conserved asparagine. Third, the alignment shows a 100% conservation of thymine in the second base positions of three codons within strand E and two codons in strand F. Apart from methionine, codons with thymine in the second position code for leucine (L), isoleucine (I), valine (V), and phenylalanine (F), all of which are hydrophobic residues (Trinquier and Sanejouand 1998) and may help to maintain the intermonomer interactions necessary for the assembly of the homohexamers required for enzyme activity in germins (see references in Dunwell, Khuri, and Gane 2000). These observations are also linked to conservation of the core region of the barrel structure of cupins, in that the length and composition of the β-strands remains largely conserved, with most variability occurring in the structurally less important central loop region between strands E and F.

Cupin ESTs

There are a rapidly increasing number of cupin EST sequences in the databases, from a wide range of species; only a sample of them was used in the phylogenies described above. An alignment of all 61 cupin

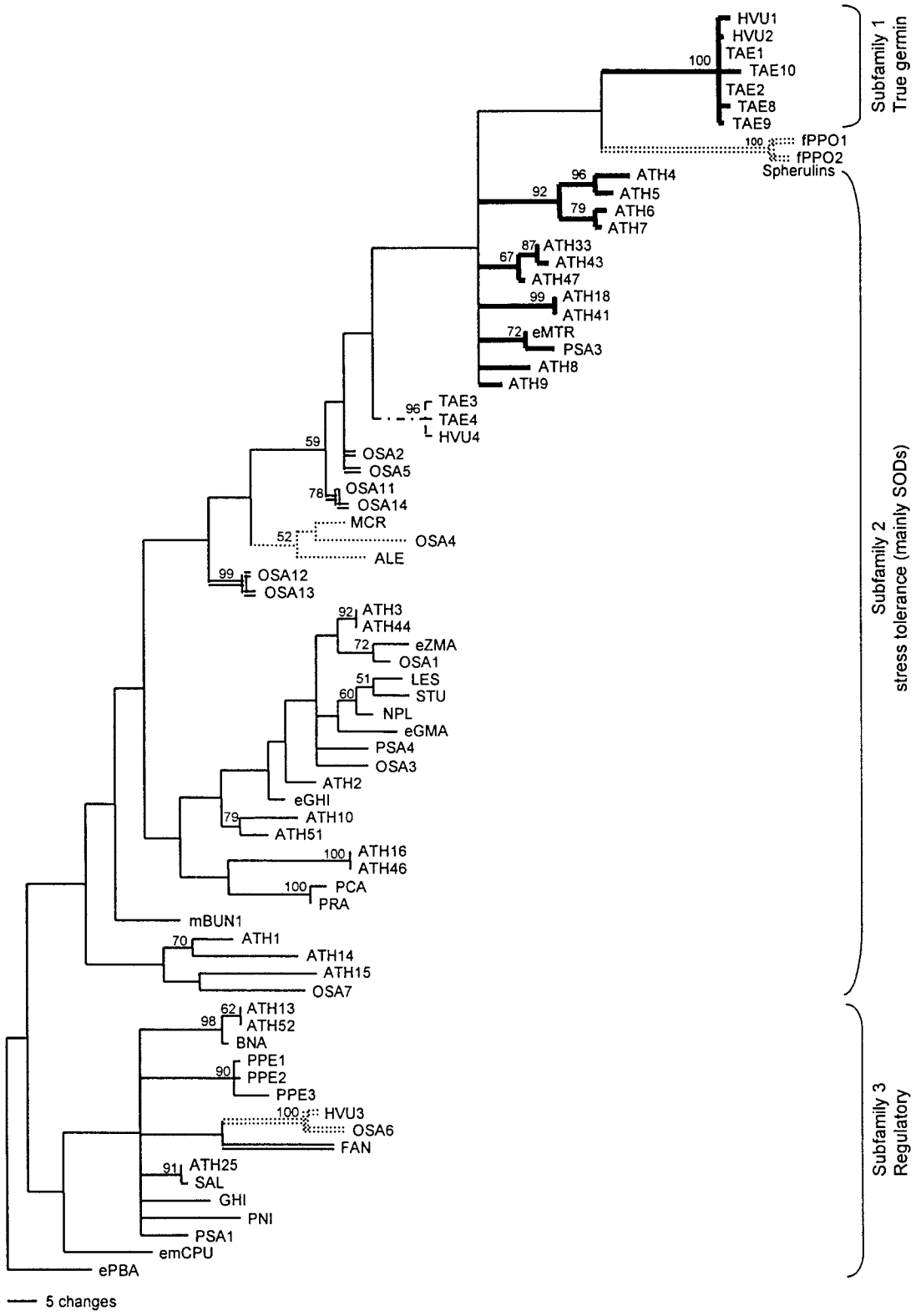

FIG. 4.—Plant and fungal GLPs: strict consensus of 2,928 most-parsimonious trees after Protpars stepmatrix character state weighting (see text), rooted on basal branch as found in the global analysis (fig. 1). Numbers at nodes indicate jackknife values (10,000 replicates). Intermotif regions are 20 amino acids, with the following exceptions: double dotted line = 21, double line = 22, bold line = 23, dotted line = 25; dot-and-dash line = 26. Prefix f = fungal; m = moss; e = EST. All sequences used and their accession numbers are presented in table 2.

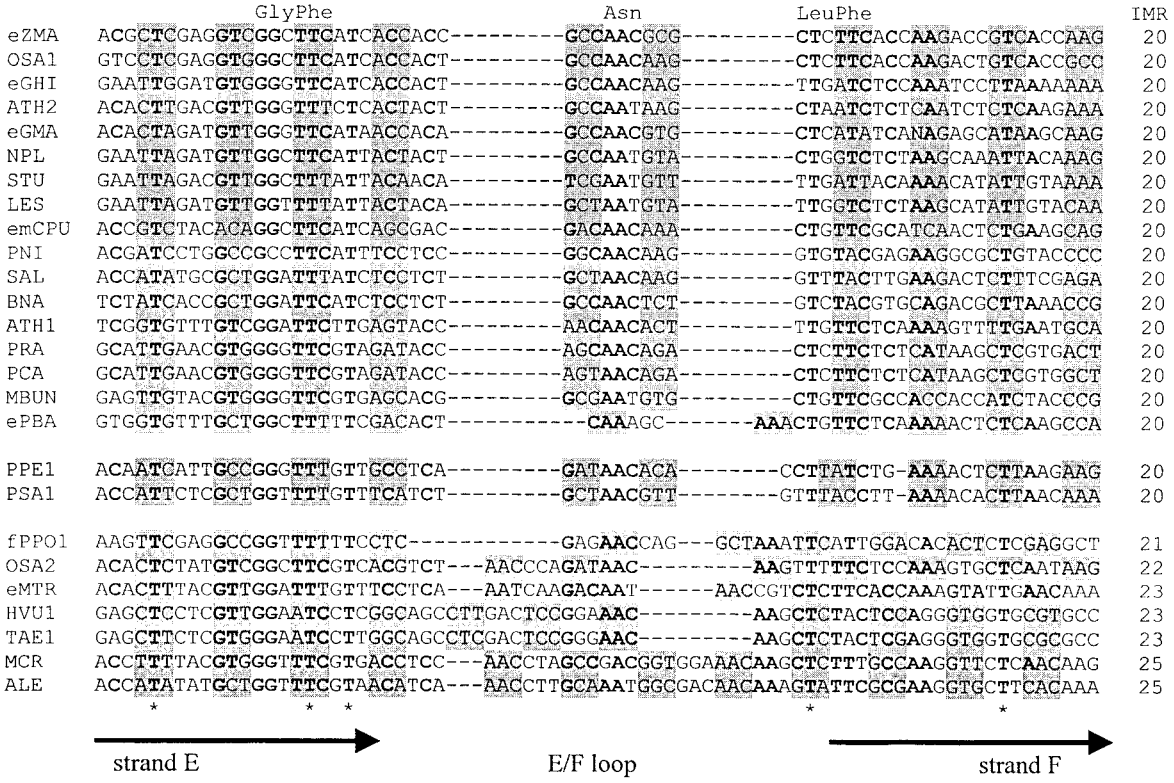

FIG. 5.—Alignment of DNA sequences encoding the intermotif regions (IMRs) of representative germin-like proteins, grouped according to the length of the IMR. Alternate codons are shaded, bases showing ≥50% conservation are in bold, and an asterisk denotes a 100% conserved thymine in the second position.

ESTs available at the time of this analysis (data not shown) confirmed the high level of conservation within the cupin motifs and the IMRs of these proteins. The cupin ESTs appeared from cDNA libraries constructed from a variety of plant tissues, such as leaf, shoot and root material, as well as callus tissue (gi|5602575) and floral tissues (gi|4152154). Furthermore, a number of these ESTs originated from tissue under biotic or abiotic stress, such as maize silks infected with Fusarium (gi|6681825) or *Medicago truncatula* roots infected with Phytophthora (gi|6654979). *Mesembryanthemum crystallinum* tissue subjected to high salt levels yields a number of cupin ESTs (gi|2911925, gi|5917863, gi|2734069, gi|5443071, gi|5442661, and gi|4464834), as does *M. truncatula* tissue grown under phosphate deficiency (gi|7675310, gi|7675658, and gi|7675990). These results are reminiscent of the expression of wheat GLPs during similar biotic (Schweizer, Christoffel, and Dudler 1999) and abiotic stress responses (Berna and Bernier 1999).

Conclusions

Cupins have been found in organisms from thermophilic bacteria to plants and animals that inhabit a whole spectrum of environments (Dunwell 1998*a*). The widespread occurrence of the conserved domain in pro- and eukaryotic proteins that include a variety of enzymes and binding proteins (Dunwell, Khuri, and Gane 2000) led to the conclusion that the compact beta-barrel structure that makes up the cupin core (Gane, Dunwell, and Warwicker 1998; Woo et al. 2000) provides a stable scaffold that allows these proteins to survive and function under a great variety of extreme conditions (Thompson and Eisenberg 1999) and can be functionally modified by relatively minor changes to the active-site region.

An unrooted phylogeny of all the known cupin motifs at the time of analysis (tables 1 and 2) gave rise to a function-based clustering (fig. 2), proving that even within the six-strand cupin domain there was sufficient information to permit discrimination between different clades. If this tree were to be rooted using the cupin (aae) from the thermophilic *A. aeolicus* (considered the most deeply branched bacterium; Deckert et al. 1998) as the outgroup, it is expected that bacterial PMIs would be the most primitive cupins, with the eukaryotic GLPs, and particularly the true germins (from cereals), being the most advanced.

In terms of functional evolution, this means that the first ''protocupin'' arose as a small β-barrel protein (about 100 residues), perhaps containing iron or manganese (Kirschvink et al. 2000) and initially able to bind a variety of simple sugars. As the environment changed, the protocupin gene was modified and diversified as the prokaryote evolved. The proteins underwent amino acid additions and/or substitutions to bind more complex sugars and fused with other ''protoproteins'' to catalyze more complex reactions. Throughout this evolutionary

process, the key structural residues were conserved and the distinctive beta-barrel structure was maintained. At some point, one such cupin fused to another, or was duplicated, with the emergence of bi-cupins such as the OXDCs. It is hoped that the data set from this superfamily of proteins will be of value in addressing the objections raised against phylogenetic analyses by those who argue that most sequences are heavily saturated with respect to amino acid substitutions and that it is usually not possible to define a stable character in a given domain (a specific conserved residue) because these positions have evolved too rapidly (Philippe and Forterre 1999). In the cupin sequences, in contrast, there are definite fixed residues known to have important structural and functional significance.

Relationship of GLPs to other Mn-SODs

Recently, two independent sources have provided evidence for a functional similarity between GLPs and the well-characterized group of iron/manganese SODs found in prokaryotes and in the mitochondria of eukaryotes. First, resolution of the three-dimensional structure of barley germin at 1.6 Å (Woo et al. 2000) revealed an active site comprising a single Mn atom liganded by the three histidines and the single glutamate within the two cupin motifs (fig. 1). This active site has close geometric similarity to that of the Mn/Fe class of SOD, which consists of three histidines and an aspartate residue (although the overall protein fold of these SODs is different from that of the cupins). Direct experimental evidence for the connection comes from the finding of SOD activity in a manganese-containing GLP from the moss *B. unguiculata* (Yamahara et al. 1999), from the detection of Mn-SOD activity from a tobacco GLP (Carter and Thornburg 2000), and from the detection of both OXO and SOD activities in a barley germin (Woo et al. 2000).

It now seems logical to suggest that the major functional role for all GLPs is as a SOD enzyme that is active at neutral pH and protects plants from oxidative stress induced by the range of biotic and abiotic stresses referred to above. Compared with the majority of plant GLPs, the more specialized cereal germins (figs. 2 and 5) are assumed to have evolved an additional OXO function at low pH. This activity provides the additional benefit of enabling detoxification of the oxalic acid produced by several plant pathogens (Dunwell, Khuri, and Gane 2000).

It is now thought that the ancestral form of the previously characterized type of microbial SOD was cambialistic (i.e., it was able to bind both Fe and Mn) and that the two isoforms (Mn form and Fe form) diverged at a later stage of evolution, after the divergence of the Archaea and Eubacteria. This change in metal binding specificity has recently been related to changes in the redox potential of the surface oceans during the periods around the intense global glaciations known as ''snowball earth'' (Kirschvink et al. 2000). The enzymatic cupin precursors may have undergone a similar evolutionary path.

As this study has shown, in this genome-sequencing era, it is possible to continue to define new structurally related families and superfamilies of functionally diverse proteins. It is imperative, however, that predictions of the enzymatic functions of proteins be supported by biochemical evidence, although this is not always a simple issue. The identification of barley germin as an OXO (Lane et al. 1993), for example, came 80 years after evidence for the enzyme was first provided (Zaleski and Reinhard 1912), and it was nearly 10 years later that its second function as a SOD was discovered (Woo et al. 2000).

Acknowledgments

The authors are grateful to the Biotechnology and Biological Sciences Research Council, United Kingdom (S.K., J.M.D.), the Natural Environment Research Council, United Kingdom (F.T.B.), and Syngenta Ltd (J.M.D.) for financial support. We also thank Dr. Alastair Culham for valuable advice, and the two reviewers for positive and constructive comments.

LITERATURE CITED

ARAVIND, L., and E. V. KOONIN. 1999*a*. Gleaning non-trivial structural, functional and evolutionary information about proteins by iterative database searches. J. Mol. Biol. **287**: 1023–1040.

———. 1999*b*. DNA-binding proteins and evolution of transcription regulation in the archaea. Nucleic Acids Res. **27**: 4658–4670.

BÄUMLEIN, H., H. BRAUN, I. A. KAKHOVSKAYA, and A. D. SHUTOV. 1995. Seed storage proteins of spermatophytes share a common ancestor with desiccation proteins of fungi. J. Mol. Evol. **41**:1070–1075.

BERNA, A., and F. BERNIER. 1999. Regulation by biotic and abiotic stress of a wheat germin gene encoding oxalate oxidase, a H_2O_2-producing enzyme. Plant Mol. Biol. **39**:539–549.

CARTER, C., and R. W. THORNBURG. 1999. Germin-like proteins: structure, phylogeny and function. J. Plant Biol. **42**: 97–108.

———. 2000. Tobacco Nectarin 1: purification and characterisation as a germin-like, manganese superoxide dismutase implicated in the defense of floral reproductive tissues. J. Biol. Chem. **275**:36726–36733.

DECKERT, G., P. V. WARREN, T. GAASTERLAND et al. (15 coauthors). 1998. The complete genome of the hyperthermophilic bacterium *Aquifex aeolicus*. Nature **392**:353–358.

DOMON, J.-M., B. DUMAS, E. LAINÉ, Y. MEYER, A. DAVID, and H. DAVID. 1995. Three glycosylated polypeptides secreted by several embryogenic cell lines of pine show highly specific serological affinity to antibodies directed against the wheat germin apoprotein monomer. Plant Physiol. **108**:141–148.

DUNWELL, J. M. 1998*a*. Cupins: a new superfamily of functionally diverse proteins that include germins and plant storage proteins. Biotechnol. Genet. Eng. Rev. **15**:1–32.

———. 1998*b*. Sequence analysis of the cupin gene family in *Synechocystis* PCC6803. Microb. Comp. Genomics **3**:141–148.

DUNWELL, J. M., and P. J. GANE. 1998. Microbial relatives of seed storage proteins; conservation of motifs in a function-

ally diverse superfamily of enzymes. J. Mol. Evol. **46**:147–154.

DUNWELL, J. M., S. KHURI, and P. J. GANE. 2000. Microbial relatives of the seed storage proteins of higher plants: conservation of structure and diversification of function during the evolution of the cupin superfamily. Microbiol. Mol. Biol. Rev. **64**:153–179.

FARRIS, J. S. 1969. A successive approximations approach to character weighting. Syst. Zool. **18**:374–385.

GALLEGOS, M.-T., R. SCHLEIF, A. BAIROCH, K. HOFMAN, and J. L. RAMOS. 1997. AraC/XylS family of transcription regulators. Microbiol. Mol. Biol. Rev. **61**:393–410.

GANE, P. J., J. M. DUNWELL, and J. WARWICKER. 1998. Modelling based on the structure of vicilins predicts a histidine cluster in the active site of oxalate oxidase. J. Mol. Evol. **46**:488–493.

HAMEL, F., C. BRETON, and M. HOUDE. 1998. Isolation and characterization of wheat aluminum-regulated genes: possible involvement of aluminum as a pathogenesis response elicitor. Planta **205**:531–538.

HEINTZEN, C., R. FISCHER, S. MELZER, K. KAPPELER, K. APEL, and D. STAIGER. 1994. Circadian oscillations of a transcript encoding a germin-like protein that is associated with cell walls in young leaves of the long-day plant *Sinapis alba* L. Plant Physiol. **106**:905–915.

HURKMAN, W. J., and C. K. TANAKA. 1996. Effect of salt stress on germin gene expression in barley roots. Plant Physiol. **110**:971–977.

KESARWANI, M., M. AZAM, K. NATARAJAN, A. MEHTA, and A. DATTA. 2000. Oxalate decarboxylase from *Collybia velutipes*. Molecular cloning and its overexpression to confer resistance to fungal infection in transgenic tobacco and tomato. J. Biol. Chem. **275**:7230–7238.

KIRSCHVINK, J. L., E. J. GAIDOS, L. E. BERTANI, N. J. BEUKES, J. GUTZMER, L. N. MAEPA, and R. E. STEINBERGER. 2000. Paleoproterozoic snowball earth: extreme climatic and geochemical global change and its biological consequences. Proc. Natl. Acad. Sci. USA **97**:1400–1405.

KOONIN, E. V., A. R. MUSHEGIAN, M. Y. GALPERIN, and D. R. WALKER. 1997. Comparison of archaeal and bacterial genomes: computer analysis of protein sequences predicts novel functions and suggests a chimeric origin for the Archaea. Mol. Microbiol. **25**:619–637.

LANE, B. G. 2000. Oxalate oxidases and differentiating surface structure in wheat: germins. Biochem. J. **349**:309–321.

LANE, B. G., F. BERNIER, E. DRATEWKA-KOS, R. SHAFAI, T. D. KENNEDY, C. PYNE, J. R. MUNRO, T. VAUGHAN, D. WALTERS, and F. ALTOMARE. 1991. Homologies between members of the germin gene family in hexaploid wheat and similarities between these wheat germins and certain *Physarum* spherulins. J. Biol. Chem. **266**:10461–10469.

LANE, B. G., J. M. DUNWELL, J. RAY, M. R. SCHMITT, and A. C. CUMING. 1993. Germin, a marker of early plant development, is an oxalate oxidase. J. Biol. Chem. **268**:12239–12242.

MADDISON, W. P., and D. R. MADDISON. 1992. MacClade. Version 3.04. Analysis of phylogeny and character evolution. Sinauer, Sunderland, Mass.

NEUTELINGS, G., J. M. DOMON, N. MEMBRE, F. BERNIER, Y. MEYER, A. DAVID, and H. DAVID. 1998. Characterization of a germin-like protein gene expressed in somatic and zygotic embryos of pine. Plant Mol. Biol. **38**:1179–1190.

ONO, M., K. SAGE-ONO, M. INOUE, H. KAMADA, and H. HARADA. 1996. Transient increase in the level of mRNA for a germin-like protein in leaves of the short-day plant *Pharbitis nil* during the photoperiodic induction of flowering. Plant Cell Physiol. **37**:855–861.

PHILIPPE, H., and P. FORTERRE. 1999. The rooting of the universal tree of life is not reliable. J. Mol. Evol. **49**:509–523.

REQUENA, L., and S. BORNEMANN. 1999. Barley (*Hordeum vulgare*) oxalate oxidase is a manganese-containing enzyme. Biochem. J. **343**:185–190.

SCELONGE, C. J., and D. L. BIDNEY. 1998. Gene encoding oxalate decarboxylase from *Aspergillus phoenices*. Patent Application WO 98/42827.

SCHWEIZER, P., A. CHRISTOFFEL, and R. DUDLER. 1999. Transient expression of members of the germin-like gene family in epidermal cells of wheat confers disease resistance. Plant J. **20**:541–552.

SHUTOV, A. D., F. R. BLATTNER, and H. BÄUMLEIN. 1999. Evolution of a conserved protein module from Archaea to plants. Trends Genet. **15**:348–349.

SOISSON, S. M., B. MACDOUGALL-SHACKLETON, R. SCHLEIF, and C. WOLBERGER. 1997. Structural basis for ligand-regulated oligomerization of AraC. Science **276**:421–425.

STAIGER, D., K. APEL, and G. TREPP. 1999. The *Atger3* promoter confers circadian clock-regulated transcription with peak expression at the beginning of the night. Plant Mol. Biol. **40**:873–882.

SWOFFORD, D. L. 1999. PAUP*: phylogenetic analysis using parsimony (*and other methods). Version 4.0b4a. Sinauer, Sunderland, Mass.

TANNER, A., and S. BORNEMANN. 2000. *Bacillus subtilis* YvrK is an acid-induced oxalate decarboxylase. J. Bacteriol. **18**: 5271–5273.

THOMPSON, M. J., and D. EISENBERG. 1999. Transproteomic evidence of a loop-deletion mechanism for enhancing protein thermostability. J. Mol. Biol. **290**:595–604.

THORDAHL-CHRISTENSEN, H., Z. ZHANG, Y. WEI, and D. B. COLLINGE. 1997. Subcellular localization of H_2O_2 in plants: H_2O_2 accumulation in papillae and hypersensitive response during powdery-mildew interaction. Plant J. **11**:1187–1194.

TRINQUIER, G., and Y. H. SANEJOUAND. 1998. Which effective property of amino acids is best preserved by the genetic code? Protein Eng. **11**:153–169.

VALLELIAN-BINDSCHEDLER, L., E. MOSINGER, J. P. METRAUX, and P. SCHWEIZER. 1998. Structure, expression and localization of a germin-like protein in barley (*Hordeum vulgare* L.) that is insolubilized in stressed leaves. Plant Mol. Biol. **37**:297–308.

WOO, E.-J., J. M. DUNWELL, P. W. GOODENOUGH, A. C. MARVIER, and R. W. PICKERSGILL. 2000. Barley germin is a manganese containing homohexamer with oxalate oxidase and superoxide dismutase activities. Nat. Struct. Biol. **7**: 1036–1040.

YAMAHARA, T., T. SHIONO, T. SUZUKI, K. TANAKA, S. TAKIO, K. SATO, S. YAMAZAKI, and T. SATOH. 1999. Isolation of a germin-like protein with manganese superoxide dismutase activity from cells of a moss, *Barbula unguiculata*. J. Biol. Chem. **274**:33274–33278.

ZALESKI, W., and A. REINHARD. 1912. Über die fermentative Oxydation der Oxalsäure. Biochem. Zeitung **33**:449–455.

ELIZABETH KELLOGG, reviewing editor

Accepted December 13, 2000