J Mol Evol (1998) 46:147–154

JOURNAL OF MOLECULAR EVOLUTION

Microbial Relatives of Seed Storage Proteins: Conservation of Motifs in a Functionally Diverse Superfamily of Enzymes

Jim M. Dunwell,[1] Paul J. Gane[2,*]

[1] Department of Agricultural Botany, School of Plant Sciences, The University of Reading, Whiteknights, PO Box 221, Reading RG6 6AS, United Kingdom
[2] Institute of Food Research, Reading Laboratory, Earley Gate, Whiteknights Road, Reading RG6 6BZ, United Kingdom

Received: 25 April 1997 / Accepted: 29 July 1997

Abstract. Plant storage proteins comprise a major part of the human diet. Sequence analysis has revealed that these proteins probably share a common ancestor with a fungal oxalate decarboxylase and/or related bacterial genes. Additionally, all these proteins share a central core sequence with several other functionally diverse enzymes and binding proteins, many of which are associated with synthesis of the extracellular matrix during sporulation/encystment. A possible prokaryotic relative of this sequence is a bacterial protein (SASP) known to bind to DNA and thereby protect spores from extreme environmental conditions. This ability to maintain cell viability during periods of dehydration in spores and seeds may relate to absolute conservation of residues involved in structure determination.

Key words: Seed storage proteins — Enzyme superfamily — Protein domain — Germin — Oxalate oxidase — Histidine cluster — Mannose metabolism

Introduction

Plant storage proteins, particularly those found in seeds, form a major part of the human diet. Many such proteins (including the vicilins, legumins, and globulins) are now known to have two domains, each similar in sequence to a family of proteins (Bäumlein et al. 1995) the best known of which is germin, the predominant protein produced during the early phase of germination of the wheat embryo (Lane et al. 1992). Germin is a glycosylated, homopentameric protein with exceptional resistance to proteases and hydrogen peroxide, the latter feature related to its function as an oxalate oxidase (Lane et al. 1993) that generates peroxide. There are several germin-like proteins found in dicotyledonous species (Heintzen et al. 1994; Ono et al. 1996) (see also the multigene family in *Arabidopsis*, accessions U75187–U75207, U95034–U95036) and gymnosperms (Domon et al. 1995). The function of these proteins is unknown; most do not have any oxalate oxidase activity, although at an amino acid level there is a high level of similarity to the cereal ox-ox enzymes (for a recent discussion see Berna and Bernier 1997).

The present study started from the specific identification (Lane et al. 1991) of the so-called ''germin box'' (HI/THPRATEI), a conserved sequence shared by the germins and spherulins—the latter are a group of proteins produced in the slime mold *Physarum polycephalum* during encystment (Bernier et al. 1987). Previous PROSITE analysis had identified at PDOC00597 a germin family signature PS00725 which comprised the germin box. Additionally, there is a three-element PRINTS fingerprint GERMIN which is based on alignment of 12 sequences and a ProDom domain 2426 (ProDom release 34.1) found in 10 proteins. In each case the analysis is

* *Present address:* The Drug Design Group, Department of Pharmacology, University of Cambridge, Tennis Court Road, Cambridge CB2 1QJ, United Kingdom
Correspondence to: J.M. Dunwell; e-mail J.Dunwell@reading.ac.uk

severely outdated since it is based on only a small subset of those sequences now available. (The SwissProt database contains only a minority of the sequences available.)

To refine these analyses, various methods were adopted in an attempt to identify possible progenitors of the germin/germin-like proteins, and also to assess the occurrence of any conserved motifs with potential active site function (see Gane et al. 1997).

Methods

Initially, BLAST searches were conducted on various germin sequences in the GenBank database, and these were linked to BLOCKS (Henikoff and Henikoff 1991), MEME (Bailey and Elkan 1994), ProDom, and PROSITE analyses. Taken together, these analyses showed (1) that only certain residues within the germin box are absolutely conserved across all proteins and (2) that other regions of the gene are equally conserved. Specifically, the germin box is part of a 20/21-AA motif $G(X)_5HXH(X)_{11}G$ (part of PRINTS motif GERMIN1) which is followed (usually after 15 residues) by a second motif of 16 AAs, $G(X)_5P(X)_4H(X)_3N$ (part of PRINTS motif GERMIN2). When one uses the bean storage protein phaseolin (Lawrence et al. 1994) and the related sucrose-binding protein (Braun et al. 1996) as structural references, these two histidine-containing motifs are part of the β-strands designated, respectively, C/D and G/H within the two β-barrel elements. Considering the first motif, the two flanking glycines correspond to the strictly conserved Gly^{249} and Gly^{269} which assist in the formation of the short interstrand loops B-C and D-E within the C-terminal β-barrel of phaseolin. Similarly, the second motif contains a proline corresponding to conserved Pro^{303} which is part of the interstrand loop between strands G and H. The variable space between motifs is equivalent to the insertions (seven to 25 residues) tolerated in the E-F loop (see Gane et al. 1997).

Gene families found to contain these two specific motifs are considered below, with summary details provided in Table 1 and Fig. 1. In order to provide a link to previous publications (Bäumlein et al. 1995; Braun et al. 1996; Lane et al. 1991; Lawrence et al. 1994) these summaries include phaseolin (gi|230247), a spherulin (P09351), a vicilin (Z54364) found in spores of the Ostrich fern *Matteuccia struthiopteris* (the C-terminal domain of this sequence is particularly similar to wheat germin) and the reference standard wheat germin gf2.8 (P15290).

It should be noted that there are two examples described below (i.e., VIRT18179 and VIRT13653) in which TBLASTN searches revealed previously unknown coding regions in upstream sequences of other genes. The explanation for the presence of these cryptic genes (sometimes encoded by sequences in different frames) probably lies in cloning artifacts caused by accidental ligation of two different DNA fragments.

Results

Oxalate Decarboxylase and Other Duplicated Proteins

This study revealed three examples of two-domain proteins (Fig. 2) in which each domain contains both of the two motifs, and there is significant similarity to GLPs and their relatives. The first, and best described in terms of function, is the oxalate decarboxylase (ODC) enzyme (I25120) (Mehta and Datta 1991) from the wood-rotting fungus *Collybia velutipes* (Fig. 2A). It is of especial interest that this oxalate-degrading enzyme should be a duplicated version of the only other enzyme with a related function—namely, the cereal oxalate oxidase described above. (However, it does not seem to be related to the oxalyl-CoA decarboxylase enzyme, gi|1086099, from *Oxalobacter formigenes*.) The closest neighbor [WU-BLASTP, P(N)3.0e-55; 37% identity and 54% similarity over 326 AA] to the *Collybia* ODC sequence is the second example, a hypothetical protein (accession D90907) from the cyanobacterium *Synechocystis* (Kaneko et al. 1996) (Fig. 2B) which is shown in a WU-BLASTP search as being most similar [P(N)2.5e-216] to the *Arabidopsis* GLP10 (U95036); it is 30% identical and 49% similar over a distance of 139 AA. The third example of duplication (accession P42106) is another protein (Yoshida et al. 1995) of unknown function, from *Bacillus subtilis*, and was identified in the ProDom study referred to later. It is one of 17 proteins containing the 52-AA domain 1428 (ProDom release 34.1) which spans the two motifs identified here, and it has as its closest neighbor [P(N)0.00028] the polyketide synthase P23157 referred to below.

Despite the fact that the critical residues within each motif have been conserved in all three examples, there are a number of significant differences between the sequences of these proteins. First, the gap between motifs in P42106 is 15 AA (the overall minimum), and 20 AA in the other two examples. Second, overlaps of the duplicated regions are of different length, and in addition, BLASTP searches with BEAUTY annotation (Worley et al. 1990) show that the stretches of sequence showing significant similarity to other proteins are located primarily in the C-terminal section of oxalate decarboxylase, whereas they are located mostly in the N-terminal half of accession P42106. The third protein, D90907, has equal similarity (particularly to germins/vicilins) in each half.

In terms of their evolution, there are a number of possible origins for this duplication, which previously has never been found outside the plant kingdom. The simplest hypothesis is that the two domains of each protein have diverged differentially after a single duplication event. Alternatively, duplication might have occurred three times, most recently in *Synechocystis* (D90907), in which the two domains show equal similarity to other proteins. The third possibility is that this latter protein is the only one to be a product of duplication, and the other two examples have evolved from a homologous recombination event between two similar DNA sequences. In this case, D90907 would be the direct and sole ancestor of the higher plant storage proteins.

Auxin-Binding Proteins

An extremely high level of similarity (66% identity; 77% similarity) was identified between two auxin-binding

Table 1. Summary of single- and double-domain proteins which contain motifs 1 and 2[a]

Species	Accession	Length (AA)	Name/function
i) Single domain			
Prokaryote:Archaea			
Desulfurococcus sp.	D84067	118	VIRT18179, unknown
Methanococcus jannaschii	U67602	125	?PMI/GDP
Prokaryote:Eubacteria:Protebacteria			
Erwinia chrysanthemi	Q05527	110	Unknown
Escherichia coli	P38522	121	Aldehyde dehydrogenase
Erwinia chrysanthemi	L39897	132	ORF 1/unknown
Enterobacter aerogenes	U60777	140	Pep1
Desulfibrio desulfuricans	Z11975	163	VIRT13653, unknown
Neisseria meningitidis	L09188	180	Deoxyglucose epimerase
Escherichia coli	gi|1788802	233	Unknown
Escherichia coli	P17410	280	Cel operon repressor
Yersinia enterocolitica	U46859	465	Mannose pyrophosphorylase
Xanthomonas campestris	P29956	466	Phosphomannose isomerase
Prokaryote:Eubacteria:Firmicutes:Actinomycete			
Streptomyces coelicolor	U37580	77	Membrane-spanning protein
Mycobacterium tuberculosis	Z81360	116	Unknown
Streptomyces cyaneus	Q02586	154	CurC/?cyclase
Mycobacterium tuberculosis	gi|1781124	263	?regulatory protein
Prokaryote:Eubacteria:Firmicutes:LowG+C gram positive			
Bacillus subtilis	gi|1881251	113	YdbB/unknown
Bacillus subtilis	P54430	186	YrkC/unknown
Acholeplasma laidlawii	S33518	369	Unknown
Staphylococcus aureus	U81973	371	Cap5F/unknown
Bacillus subtilis	P39631	432	Spore polysacch. synth.
Clostridium thermocellum	P26208	448	β-glucosidase A
Prokaryote:Eubacteria:Cyanobacteria			
Synechocystis sp. PCC6803	D64001	128	Phosphomannose isomerase
Synechocystis sp. PCC6803	D90909	135	Unknown
Synechocystis sp. PCC6803	D90910	143	Unknown/?cytochrome c551
Synechococcus sp.	S04426	150	Unknown/?PMI
Eukaryote:Fungi:Ascomycetes			
Saccharomyces cerevisiae	P47096	177	?3-HAO
Saccharomyces cerevisiae	S53039	179	Unknown
Eukaryote:Myxomycetes			
Physarum polycephalum	P09351	248	Spherulin
Eukaryote:Plant:Angiosperm			
Arabidopsis thaliana	P33487	198	Auxin-binding protein
Triticum aestivum	P15290	201	Germin/oxalate oxidase
Prunus persica	gi|1916807	214	Auxin-binding protein
Arabidopsis thaliana	Q05212	230	DRT 102/DNA damage repair
Vicia faba	X95995	1641	ENBP1/zinc finger protein
Eukaryote:Animal			
Caenorhabditis elegans	gi|1082118	104	Unknown
Caenorhabditis elegans	gi|1707132	159	Unknown
Caenorhabditis elegans	P39645	190	RFBC/epimerase
Caenorhabditis elegans	Z70755	207	3-HAO
Caenorhabditis elegans	gi|726397	221	Unknown
Caenorhabditis elegans	gi|1572766	349	Unknown
Caenorhabditis elegans	Z50070	645	Unknown
Caenorhabditis elegans	gi|1572765	910	Unknown
Caenorhabditis elegans	U00043	1070	Unknown
Rattus norvegicus	P21816	200	Cysteine dioxygenase
Rattus rattus	D44494	286	3-HAO
Rattus norvegicus	X59993	1214	Zinc finger protein
ii) Double domain			
Prokaryote:Eubacteria:Firmicutes:LowG+C gram positive			
Bacillus subtilis	P42106	337	Unknown
Prokaryote:Eubacteria:Cyanobacteria			
Synechocystis sp.PCC6803	D90907	394	Unknown
Eukaryote:Fungi:Basidiomycete			
Collybia velutipes	I25120	447	Oxalate decarboxylase
Eukaryote:Plant:Pteridophyte			
Matteuccia struthiopteris	Z54364	504	Spore vicilin
Eukaryote:Plant:Angiosperm			
Phaseolus vulgaris	gi|230247	397	Phaseolin

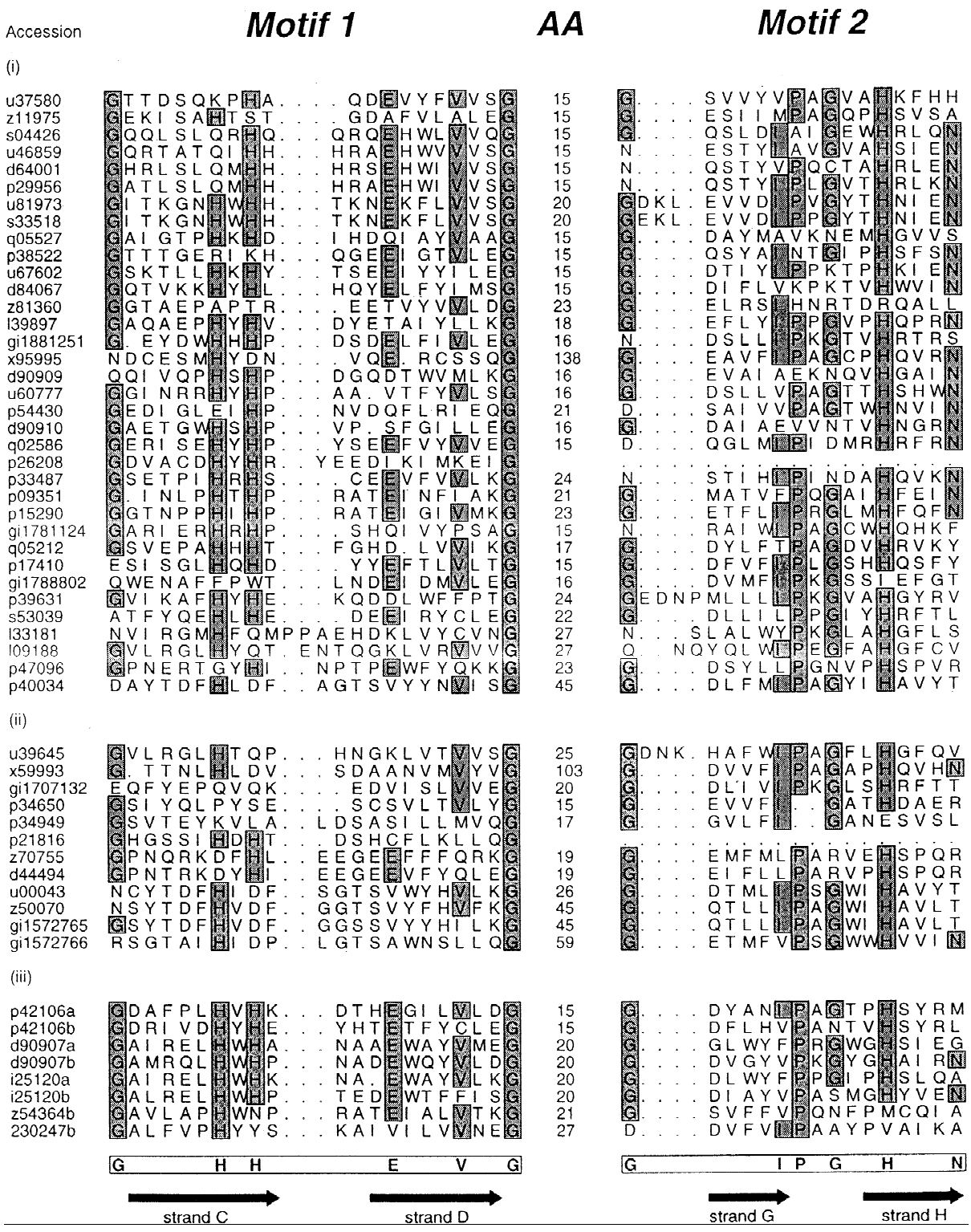

Fig. 1. Details of the two conserved motifs and their spacing, present in a range of proteins identified by their GenBank accession numbers. Section *(i)* denotes microbial and plant proteins, section, *(ii)* animal proteins, and section *(iii)* two-domain proteins. Suffixes *(a)* and *(b)* represent, respectively, the first and second domains in these two-domain proteins. Highly conserved residues are shaded and are shown in the boxes below each alignment. Nomenclature of β-strands is according to Lawrence et al. (1994).

proteins (ABPs) from *Prunus persica* (gi|1916807, gi|1916809) and several germins (e.g., U75193) from *Arabidopsis thaliana.* Considering only the two motifs, this similarity to U75193 increases to 17/21 (81%) identity for motif 1 and 14/16 (88%) identity for motif 2. Indeed, on the basis of their sequence, these peach ABPs should probably be reclassified as germin-like proteins (GLPs). However, extensive similarities to GLPs were also detected in sequences from functionally defined ABPs (Venis and Napier 1995) such as that from *A.*

a). Oxalate decarboxylase (I25120) from *Collybia velutipes*

```
   1 MFNNFQRLLTVILLSGFTAGVPLASTTTGTGTATGTSTAAEPSATVPFASTDPNPVLWNETSDPALVKPE
 227                                              TFLLTDWLSHVPMEVALKNFRAKNP

  71 RNQLGATIQGPDNLPADLQNPDLLAPPTTDHGFVGNAKWPFSFSKQRLQTGGWARQQNEVVLPLATNLAC
 252 AAWSHIPAQQLYIFPSEPPADNQPDPVSPQGTVPLPYSFNFSSVEPTQYSGGTAKIADSTTFNISVAIAV
                              *              **        ** *              *

 141 TNMRLEAGAIRELHWHKNA-EWAYVLKGSTQISAVDNEGRNYISTVGPGDLWYFPPGIPHSLQATADDPE
 322 AEVTVEPGALRELHWHPTEDEWTFFISGNARVTIFAAQSVASTFDYQGGDIAYVPASMGHYVENIGNTTL
          * ** ******    **     *                    **    *    *

 211 GSEFILVFDSGAFNDDGTFLLTDWLSHVPMEVALKNFRAKNPAAWSHIPAQQLYIFPS 267
 392 TYLE--VFNTDRFADVSLSQWLALTPPSVVQAHLNLDDETLAELKQFATKATVVGPVN 447
           **    * *
```

b). Hypothetical protein (D90907) from *Synechocystis*

```
   1 MVNSVIGWLRRRFLLVGLSVLLITFLGIFTPTIAQSEQWRSLSNVVWGKDLPAFTYAFSKTPLVLYDGGT
 183 TPIAWVEENLGWTAAQVAQLPKKQVYISSYGPASGPLASATPQGQTAKIEVPHTHNLLGQQPLVSLGGNE
                                                        *         ***   *

  71 TKQVGTYNFPVSKGMAGVYMSLEPGAIRELHWHANAAEWAYVMEGRTRITLTSPEGKVEIADVDKGGLWY
 253 LRLASAKEFPGSFNMTGALIHLEPGAMRQLHWHPNADEWQYVLDGEMDLTVFASEGKASVSRLQQGDVGY
             ** *  * *    ***** * **** ** ** **  *    *    ***        *

 141 FPRGWGHSIEGIGPDTAKFLLVFNDGTFSEGATFSVTDWLSHTPIAWVEENLGWTAAQVAQLPKKQVYIS
 323 VPKGYGHAIRNSSQKPLDIVVVFNDGDYQSIDLSTWLASNPSSVLGNTFQISPELTKKLPVQDTIFSLPT
      * * ** *            *****

 211 SY 212
 393 QP 394
```

c). Hypothetical protein (P42106) from *Bacillus subtilis*

```
   1 MKTLCTHSLPKEKMPYLLRSGEGERYLFGRQVATVMANGRSTGDLFEIVLLSGGKGDAFPLHVHKDTHEG
 173 AKLAELTELPDGAVPYVLESGEGDRLLTGDQLHRIVAAQKNTDGQFIVVSSEGPKGDRIVDHYHEYHTET
      *      **    ** * **** * * * *     *    *   *  *   * ***    * *    *

  71 ILVLDGKLELTLDGERYLLISGDYANIPAGTPHSYRMQSHRTRLVSYTMKGNVAHLYSVIGNPYDHAEHP
 243 FYCLEGQMTMWTDGQEIQLNPGDFLHVPANTVHSYRLDSHYTKMVGVLVPGLFEPFFRTLGDPYEGHIFP
      * * *      **    *  **    *  * ****  ** *  *     *         * **     *

 141 PYASEEVSNERFAEAAAVATIVFLDEAKPACS 172
 313 CKPQALRFDRILQNIEALDLKVMKP        337
```

Fig. 2. Sequence comparisons showing the two-domain nature of three microbial proteins—namely oxalate decarboxylase from *Collybia velutipes* **(a)** and hypothetical proteins from *Synechocystis* **(b)** and *Bacillus subtilis* **(c)**. The *underlined* residues represent intervening sequences between the duplicated segments and the *blocked regions* denote the two conserved motifs. *Asterisks* denote identity of residues within the two duplicated domains.

thaliana (P33487) (Fig. 1). Interestingly, the ABP oligopeptide designated D16 (Venis and Napier 1995), or Box 1 (Brown and Jones 1994), and thought to be involved in the binding of the carboxylic acid group, contains the $HXH(X)_4E$ (equivalent to the germin box). In terms of function, there is an interesting potential link between germin, GLPs, and ABPs in that the wheat germin promoter contains several auxin-responsive elements (Berna

and Bernier 1997). It is therefore possible that an ABP could modify the expression of germin by reducing the level of free auxin.

Polyketide Synthases

Amongst the sequences with the closest similarity [$G(X)_5HYHPYSEE(X)_6G$] to GLPs is accession M33704. This *Streptomyces cyaneus* gene (Bergh and Uhlen 1992), *curC* (probably a cyclase), is part of the synthetic pathway of the antibiotic curamycin. Closely related genes in other *Streptomyces* species include P23157, Q05362, and P16558 (see Fig. 4 in Blanco et al. 1993, which also identified residues 47–60, PGERISEHYHPYSE, cf. motif 1, as being conserved amongst several β-glucosidases; see P26208 from *Clostridium thermocellum*). Additional members include the *B. subtilis* sequence P54430 and the sequence Pep1 (Smith et al. 1993) (U60777) from *Enterobacter aerogenes.* The smallest (110–132 AA) proteins in this group are those from *Erwinia chrysanthemi* (accessions Q05527, L39897) and *B. subtilis* (gi|1881251).

Phosphomannose Isomerase and Related Enzymes Concerned With Polysaccharide Synthesis

Phosphomannose isomerases (PMIs) are zinc-containing enzymes that catalyze the interconversion of mannose-6-phosphate and fructose-6-phosphate. The class II PMIs (Proudfoot et al. 1994) are involved in a variety of pathways including capsular polysaccharide biosynthesis and D-mannose metabolism. Their similarity to proteins of the previous section was already known from the ProDom database, which includes 17 proteins with the domain 1428. These 17 include PMIs, polyketide synthases, and the duplicated *B. subtilis* sequence referred to above. Notable amongst the former category are the spore polysaccharide protein (P39631) from *B. subtilis,* the Cap5F and Cap8F proteins from *Staphylococcus aureus* (Sau and Lee 1996), and the closely related (>90%) sequence (S33518) from the mycoplasma *Acholeplasma laidlawii.* This class of PMI includes, toward their C-terminus, histidine-containing motifs similar in sequence and spacing to those described above. One difference is that the histidines of the first motif are often displaced to give a $G(X)_7HXH(X)_9G$ sequence. Of this group of proteins, the smallest (128 AA) is from *Synechocystis* (D64001). Others include: a sequence from *Synechococcus* (S04426); a sequence (created as Swissprot sequence VIRT18179) in the upstream region of the aspartate racemase gene (D84067) from the archaeon *Desulfurococcus* (Yohda et al. 1996); GDP-mannose pyrophosphorylase (Q46859) from *Yersinia enterocolitica;* and P07874—the 56-kDa bifunctional enzyme from *Pseudomonas aeruginea* with both PMI and GDP-mannose pyrophosphorylase activities (May et al. 1994). This latter enzyme is involved in the polymerization of alginate—a compound that protects the cell from host immune responses and antibiotics and is also the major cause of mortality in patients suffering from cystic fibrosis.

There are also eukaryotic equivalents of these enzymes (e.g., S53039 from yeast) in which either one or both motifs are conserved. Related sequences from *Caenorhabditis elegans* include two with unknown function, gi|1707132 and gi|1082118.

Epimerases Involved in Cell Wall Synthesis

In addition to those of the previous section, another group of capsule enzymes share the two-motif structure. These are the epimerases such as TDP-deoxyglucose epimerase (L09188) from *Neisseria meningitidis* and a similar sequence (L33181) from *Yersinia pseudotuberculosis* (Thorson et al. 1994). A smaller member of this family is the sequence (Z81360) recently reported from *Mycobacterium tuberculosis.* This study also identified one sequence, of animal origin, with particular similarity to the *Neisseria* epimerase described above. This is a DTDP-4-rhamnose-3,5-epimerase (U39645) from *C. elegans.*

Eukaryotic Dioxygenases

One class of these iron-containing enzymes contains the two motifs identified here, although the first motif has only a single histidine. These enzymes are the 3-hydroxyanthranilate-3,4-dioxygenases (3-HAO) that catalyze the synthesis of quinolinic acid from 3-hydroxyanthranilic acid. They include sequences from yeast (P47096), *C. elegans* (Z70755), and rat (D44494) (Malherbe et al. 1994). Additionally, the cysteine dioxygenases such as that (P21816) from rat (Hosokawa et al. 1990) clearly contain an equivalent first motif (Fig. 1), although the second is either absent or present as a weakly similar motif separated by a gap of approximately 40 residues.

Proteins Related to DNA Structure and Metabolism

Use of the consensus of domain 1428 in a BLASTP search shows other nonenzymatic proteins which contain the two motifs. One such example is the sequence (X52890) encoding the 280-AA CelD operon repressor from *E. coli* (Parker and Hall 1990). This protein is a unique member of the ARAC/XYLS family of transcriptional regulators (Bustos and Schleif 1993); exceptionally, it is a repressor protein, whereas all other members are positive regulators. With the exception of the recently identified sequences gi|1781124 from *M. tuberculosis* and gi|1788802 from *E. coli,* the other members of the family contain only the second motif. Three other DNA-binding proteins known to contain only the second motif are zinc-finger proteins from the rat (X59993) (Hoog et

al. 1991), mouse (Z32675), bean (X95995), and *Arabidopsis* (gi|1922960). They all contain this motif about 70 residues from the C-terminus.

Also in this category is an *Arabidopsis* protein (DRT 102) concerned with DNA damage repair/toleration and identified in a search for genes to complement *E. coli* mutants lacking defence against UV-light damage to DNA (Pang et al. 1993). Such a function may be relevant to the discussion on the SASP proteins below.

SASPs and Other Small Proteins

Additional analysis identified a series of progressively smaller proteins, each of which contains motifs 1 and 2. Amongst these proteins, none of which has any known function, are two similar accessions (D90909,D90910) from *Synechocystis*, the first of which has the *Arabidopsis* GLP X91957 as the closest relative, and the second of which is similar to cytochrome c551 from *Rhodococcus.* Slightly shorter sequences include U67602 from predicted coding region MJ1618 of *Methanococcus jannaschii* (Bult et al. 1996) and two *E. coli* sequences, (P38522) similar to an aldehyde dehydrogenase (Heim and Strehler 1991), and its longer version—the immunity repressor protein (D90768). The smallest (77 AA) of all those identified is a membrane-spanning protein (MSP) (Li and Strohl 1996) (U37580) from *Streptomyces coelicolor*. This protein, which has similarity to *E. coli* PMI, and also to an upstream sequence (SwissProt virtual sequence VIRT13653) in the prismane gene (Z11975) from *Desulfovibrio desulfuricans* (Stokkermans et al. 1992), is the putative progenitor of all these two motif proteins.

Subsequently, a BLASTP search revealed similarity (11/29 residues identical) between a region spanning the central (intermotif) portion of the MSP sequence and part of the 71-AA small, acid-soluble spore protein (SASP) from *Thermactinomyces thalpophilus* (M13042). This is one member of a family of proteins that bind to spore DNA (double stranded) and cause the DNA to change to an A-like conformation. They thus protect the DNA backbone from enzymic and nonenzymic cleavage (Setlow 1995). Of particular relevance is their resistance to hydrogen peroxide (Popham et al. 1995)—a most unusual feature also found in the cereal oxalate oxidase described above.

Discussion

There are two main conclusions to be drawn from this study. First, it has identified for the first time a series of prokaryotic and eukaryotic microbial proteins (Fig. 2) with strong similarity to the two-domain structure of the storage proteins of higher plants. Whether any oxalate decarboxylase (cf. I25120) or other enzyme activity has been retained during this evolution is unknown. Second, it provides evidence that amongst a broad range of mostly microbial organisms, there has been evolution, presumably by recombination and minor duplication, to produce several progressively larger proteins, ranging in length from the 77-AA *Streptomyces* membrane-spanning protein to the 1,214-AA rat zinc-finger protein. Each protein contains a conserved sequence, usually including two histidine-containing motifs separated by 15–27 residues (summarized in Table 1); the exact distance between motifs is a diagnostic feature of the specific class of protein. It is known that homologous proteins can evolve either different (Murzin 1993) or related (Babbit et al. 1995) enzymatic activities, but this series is possibly unique in the range of function ascribed to the various proteins—namely, types of isomerase, epimerase, cyclase, oxidase, dioxygenase, decarboxylase, and dehydrogenase, as well as binding proteins for auxin, sucrose, and DNA. Many of these enzymes are involved in production of the extracellular matrix, particularly as part of the sporulation/encystment process. This observation is in agreement with the results of a much more limited study on *E. coli* (Labedan and Riley 1995) in which most of the sequence-related proteins were also related in cellular function; they concluded that 971 paralogous genes could have been derived from only 204 ancestral genes.

Preliminary predictions of tertiary structure of the smaller single-domain proteins described in this study reveal a predominantly β-strand form (data not shown). Presumably, the terminal α-helices found in the larger proteins (>ca. 150 AA), along with the extended gap (>15 AA) between motifs, were added to the core sequence at a very early stage of evolution. These additions to the ends and to the center of the sequence occurred prior to the development of land plants and were part of the increasing complexity that led eventually to the trimeric structure of the desiccation tolerant storage proteins (approx. molecular mass 150–200 kDa) found in seeds. It can be anticipated (see Gane et al. 1997) that structural studies will soon help to describe the exact role of the two motifs, determine the components which provide the resistance to environmental extremes, and identify the various catalytic residues.

Acknowledgments. J.M. Dunwell should like to thank the BBSRC and ZENECA plc. for financial support during the course of this study.

References

Babbitt PC, Mrachko GT, Hasson MS, Huisman GW, Kolter R, Ringe D, Petsko GA, Kenyon GL, Gerlt JA (1995) A functionally diverse enzyme superfamily that abstracts the α protons of carboxylic acids. Science 267:1159–1161

Bailey TL, Elkan C (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc 2nd Int Conf Intell Systems for Mol Biol. AAAI Press, Menlo Park, California, pp 28–36

Bäumlein H, Braun H, Kakhovskaya IA, Shutov AD (1995) Seed stor-

age proteins of spermatophytes share a common ancestor with desiccation proteins of fungi. J Mol Evol 41:1070–1075

Bergh S, Uhlen M (1992) Analysis of a polyketide synthesis-encoding gene cluster of *Streptomyces curacoi.* Gene 117:131–136

Berna A, Bernier F (1997) Regulated expression of a wheat germin gene in tobacco: oxalate oxidase activity and apoplastic localization of the heterologous protein. Plant Mol Biol 33:417–429

Bernier F, Lemieux G, Pallotta D (1987) Gene families encode the major encystment-specific proteins of *Physarum polycephalum* plasmodia. Gene 59:265–277

Blanco G, Brian P, Pereda A, Méndez C, Salas JA, Chater KF (1993) Hybridization and DNA sequence analyses suggest an early evolutionary divergence of related biosynthetic gene sets encoding polyketide antibiotics and spore pigments in *Streptomyces* spp. Gene 130:107–116

Braun H, Czihal A, Shotov AD, Baumlein H (1996) A vicilin-like seed protein of cycads: similarity to sucrose-binding proteins. Plant Mol Biol 31:35–44

Brown JC, Jones AM (1994) Mapping the auxin-binding site of auxin-binding protein 1. J Biol Chem 269:21136–21140

Bult CJ et al. (1996) Complete genome sequence of the methanogenic archaeon *Methanococcus jannaschii.* Science 273:1058–1073

Bustos SA, Schleif RF (1993) Functional domains of the AraC protein. Proc Natl Acad Sci 90:5638–5642

Domon J-M, Dumas B, Lainé E, Meyer Y, David A, David H (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

Gane PJ, Dunwell JM, Warwicker J (1997) Modelling based on the structure of vicilins predicts a histidine cluster in the active site of oxalate oxidase. J Mol Evol (in press)

Heim R, Strehler EE (1991) Cloning an *Escherichia coli* gene encoding a protein remarkably similar to mammalian aldehyde dehydrogenases. Gene 99:15–23

Heintzen C, Fischer R, Melzer S, Kappeler S, Apel K, Staiger D (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

Henikoff S, Henikoff JG (1991) Automated assembly of protein blocks for database searching. Nucleic Acids Res 19:6565–6572

Hoog C, Schalling M, Grunder-Brundell E, Daneholt B (1991) Analysis of a murine germ cell-specific transcript encodes a putative zinc finger protein. Mol Reprod Dev 30:173–181

Hosokawa Y, Matsumoto A, Oka J, Itakura H, Yamaguchi K (1990) Isolation and characterization of a cDNA for rat liver cysteine dioxygenase. Biochem Biophys Res Commun 168:473–478

Kaneko T et al. (1996) Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. DNA Res 3:109–136

Labedan B, Riley M (1995) Gene products of *Escherichia coli:* sequence comparisons and common ancestries. Mol Biol Evol 12: 980–9878

Lane BG, Bernier F, Dratewka-Kos E, Shafei R, Kennedy TD, Pyne C, Munro JR, Vaughan T, Walters D, Altomare F (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 266:10461–10468

Lane BG, Cuming AC, Fregeau J, Carpita NC, Hurkman WJ, Bernier F, Dratewka-Kos E, Kennedy TD (1992) Germin isoforms are discrete temporal markers of early plant development. Eur J Biochem 209:961–969

Lane BG, Dunwell JM, Ray JA, Schmitt MR, Cuming AC (1993) Germin, a protein marker of early plant development, is an oxalate oxidase. J Biol Chem 268:12239–12242

Lawrence MC, Izard T, Beuchat M, Blagrove RJ, Colman PM (1994) Structure of phaseolin at 2.2 Å resolution: Implications for a common vicilin/legumin structure and the genetic engineering of seed storage proteins. J Mol Biol 238:748–776

Li Y, Strohl WR (1996) Cloning, purification, and properties of a phosphotyrosine protein from *Streptomyces coelicolor* A3(2). J Bacteriol 178:136–142

Malherbe P, Kohler C, Da Prada M, Lang G, Kiefer V, Schwarcz R, Lahm HW, Cesura AM (1994) Molecular cloning and functional expression of human 3-hydroxyanthranilic-acid dioxygenase. J Biol Chem 269:13792–13797

May TB, Shinabarger D, Boyd A, Chakrabarty AM (1994) Identification of amino acid residues involved in the activity of phosphomannose isomerase-guanosine 5′-diphospho-D-mannose pyrophosphorylase. A bifunctional enzyme in the alginate biosynthetic pathway of *Pseudomonas aeruginosa.* J Biol Chem 269:4872–4877

Mehta A, Datta A (1991) Oxalate decarboxylase from *Collybia velutipes.* Purification, characterisation and cloning. J Biol Chem 266: 23548–23553

Murzin AG (1993) Can homologous proteins evolve different enzymatic activities? Trends Biochem Sci 18:403–405

Ono M, Sage-Ono K, Inoue M, Kamada H, Harada H (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

Pang Q, Hays JB, Rajagopal I, Schaefer TS (1993) Selection of Arabidopsis cDNAs that partially correct phenotypes of *Escherichia coli* DNA-damage-sensitive mutants and analysis of two plant cDNAs that appear to express UV-specific dark repair activities. Plant Mol Biol 22:411–426

Parker LL, Hall BG (1990) Characterisation and nucleotide sequence of the cryptic *cel* operon of *Escherichia coli* K12. Genetics 124:455–471

Popham DL, Sengupta S, Setlow P (1995) Heat, hydrogen peroxide, and UV resistance of *Bacillus subtilis* spores with increased core water content and with or without major DNA-binding proteins. Appl Environ Microbiol 61:3633–3638

Proudfoot AEI, Turcatti G, Wells TNC, Payton MA, Smith DJ (1994) Purification, cDNA cloning and heterologous expression of human phosphomannose isomerase. Eur J Biochem 219:415–423

Sau S, Lee CY (1996) Cloning of type 8 capsule genes and analysis of gene clusters of different capsular polysaccharides in *Staphylococcus aureus.* J Bacteriol 178:2118–2126

Setlow P (1995) Mechanisms for the prevention of damage to DNA in spores of *Bacillus* species. Ann Rev Microbiol 49:29–54

Smith CA, Pinkney M, Guiney DG, Thomas CM (1993) The ancestral IncP replication system consisted of contiguous *oriV* and *trfA* segments as deduced from a comparison of the nucleotide sequences of diverse IncP plasmids. J Gen Microbiol 139:1761–1766

Stokkermans JPWG, Pierik AJ, Wolbert RBG, Hagen WR, Van Dongen WMAM, Veeger C (1992) The primary structure of a protein containing a putative [6Fe-6S] prismane cluster from *Desulfovibrio vulgaris* (Hildenborough). Eur J Biochem 208:435–442

Thorson JS, Lo SF, Ploux O, He X, Liu HW (1994) Studies on the biosynthesis of 3,6-dideoxyhexoses: molecular cloning and characterization of the asc (ascarylose) region from *Yersinia pseudotuberculosis* serogroup VA. J Bacteriol 176:5483–5493

Venis MA, Napier R (1995) Auxin receptors and auxin binding proteins. Crit Rev Plant Sci 14:27–47

Worley KC, Wiese BA, Smith RF (1990) BEAUTY: an enhanced BLAST-based search tool that integrates multiple biological information resources into sequence similarity search results. Genome Res 5:173–184

Yohda M, Endo I, Abe Y, Ohta T, Iida T, Maruyama T, Kagawa Y (1996) Gene for aspartate racemase from the sulfur-dependent hyperthermophilic archaeum, *Desulfurococcus* strain SY. J Biol Chem 271:22017–22021

Yoshida K, Seki S, Fujimura M, Miwa Y, Fujita Y (1995) Cloning and sequencing of a 36-kb region of the *Bacillus subtilis* genome between the *gnt* and *iol* operons. DNA Res 2:61–69