MICROBIAL & COMPARATIVE GENOMICS
Volume 3, Number 2, 1998
Mary Ann Liebert, Inc.

Sequence Analysis of the Cupin Gene Family in *Synechocystis* PCC6803

JIM M. DUNWELL

ABSTRACT

The recently described cupin superfamily of proteins includes the germin and germinlike proteins, of which the cereal oxalate oxidase is the best characterized. This superfamily also includes seed storage proteins, in addition to several microbial enzymes and proteins with unknown function. All these proteins are characterized by the conservation of two central motifs, usually containing two or three histidine residues presumed to be involved with metal binding in the catalytic active site. The present study on the coding regions of *Synechocystis* PCC6803 identifies a previously unknown group of 12 related cupins, each containing the characteristic two-motif signature. This group comprises 11 single-domain proteins, ranging in length from 104 to 289 residues, and includes two phosphomannose isomerases and two epimerases involved in cell wall synthesis, a member of the pirin group of nuclear proteins, a possible transcriptional regulator, and a close relative of a cytochrome c551 from *Rhodococcus*. Additionally, there is a duplicated, two-domain protein that has close similarity to an oxalate decarboxylase from the fungus *Collybia velutipes* and that is a putative progenitor of the storage proteins of land plants.

INTRODUCTION

The cupin superfamily (Dunwell, 1998) of functionally diverse proteins has been designated recently to include the germin and germinlike proteins from plants, their duplicated two-domain relatives, including the seed storage proteins (Bäumlein et al., 1995), and a wide range of other enzymes and binding proteins from microbes, plants, and animals (Dunwell and Gane, 1998). The name cupin (from the Latin *cupa*, a small barrel or cask) is derived from the tertiary β-barrel element, which comprises either the central core of these proteins (e.g., oxalate oxidase) or one of a number of discrete domains (e.g., araC transcription factors). Characteristically, the cupin element of these proteins has two histidine-containing motifs, which together with other conserved proline and glycine residues make up the structural framework and the putative metal-binding active site (Gane et al., 1998). The two conserved motifs are separated by a variable region, usually 15–20 residues in length.

The aim of the present study was to identify and categorize all cupin sequences within a single bacterial genome, namely, that of the unicellular cyanobacterium *Synechocystis* PCC6803 (Kaneko et al., 1996). This organism serves as the prokaryotic model for studying plantlike oxygenic photosynthesis, and the intention of the present study was to provide a basis from which the proliferation of related plant cupins could

Department of Agricultural Botany, School of Plant Sciences, The University of Reading, Reading RG6 6AS, UK.

be examined systematically. For example, it is estimated that the *Arabidopsis* genome contains at least 12 germinlike proteins (GLPs) (Dunwell, 1998), in addition to a large number of other related cupins.

METHODS AND RESULTS

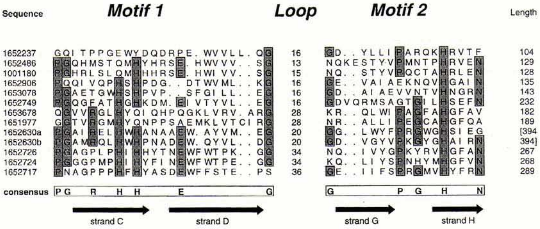
A series of detailed database searches was conducted using the gapped BLAST (Altschul et al., 1997) and BLOCKS (Henikoff and Henikoff, 1991) programs to identify those *Synechocystis* protein sequences that contain the two conserved histidine-containing motifs described by Dunwell and Gane (1998). These two motifs are part of the β-strands designated, respectively, C/D and G/H within the two β-barrel elements of the bean storage protein phaseolin (Lawrence et al., 1994). A major theme in the present analysis, and the probable reason that this gene family had not been identified previously, is that the region between the two motifs is variable in length, with a minimum distance of 15 residues for many of the bacterial proteins, increasing to around 20 for the germins and GLPs and >20 for the storage proteins. The maximum of 64 residues is found in a barley globulin (gi|421978). This variable region, which can tolerate a range of insertions, is equivalent to the D/F loop of the β-barrel structure.

The main result achieved in the present analysis was identification of a total of 12 cupin sequences, of which 11 are single-domain proteins, with one example of a two-domain structure (Fig. 1). Each of these 12 has the characteristic cupin two-motif arrangement, with the consensus of motif 1 being PG$(X)_5$HXH$(X)_4$E$(X)_7$G and that of motif 2 being G$(X)_5$PXG$(X)_2$H$(X)_3$N.

The sequences can be subdivided into several classes on the basis of a range of criteria and an analysis of their nearest neighbors according to a BLASTP search (Table 1). In an assessment of potential function, and considering first the 11 single-domain proteins, two of the sequences (gi|1001180, gi|1652486) are thought to be phosphomannose isomerases (PMIs), one (gi|1653678) to be a dTDP-4-dehydrorhamnose 3,5-epimerase, and one (gi|1651977) a dTDP-6-deoxy-L-mannose-dehydrogenase. The latter two are very similar in sequence and should probably both be considered as epimerases.

It should be noted that residues 61–129 of the PMI sequence gi|1652486 are identical to the sequence encoded by nucleotides 3–208 (with a frameshift correction at position 104) in the upstream region of gi|287460, a sequence including the *Synechocystis groES* and *groEL* genes (Lehel et al., 1993). Presumably, this partial PMI coding region was accidentally ligated to the other coding regions during the cloning procedure and then was not identified because of the frameshift introduced by a sequencing error.

Another sequence from which a function can be deduced with some reliability is gi|1652717. This has as its closest neighbor the *Escherichia coli* sequence gi|1176281, a member of the recently designated group



Sequence	Motif 1	Loop	Motif 2	Length
1652237	G Q I T P P G E W Y D Q D R P E . W V V L L . Q G	16	G D . . Y L L I P A R Q K H R V T F	104
1652486	P G Q H M S T Q M H Y H R S E . H W V V V . . S G	13	N Q K E S T Y V P M N T P H R V E N	129
1001180	P G H R L S L Q M H H H R S E . H W I V V . . S G	15	N Q . . S T Y V P Q C T A H R L E N	128
1652906	P Q Q I V Q P H S H P D G . . D T W V M L . . K G	16	G E . . V A I A E K N Q V H G A I N	135
1653078	P G A E T G W H S H P V P . . S F G I L L . . E G	16	G D . . A I A E V V N T V H N G R N	143
1652749	P G Q G F A T H G H K D M . E I V T Y V L . . E G	16	G D V Q R M S A G T G I L H S E F N	232
1653678	Q G V V R G L H Y Q I Q H P Q G K L V R V . A R G	28	K R . . Q L W I P A G F A H G F A V	182
1651977	G G T V R G M H Y Q N P P S A E M K L V T C I R G	28	N R . . A L L I P E G C A H G F Q A	189
1652630a	P G A I R E L H W H A N A A E W . A Y V M . . E G	20	G G . . L W Y F P R G W G H S I E G	[394
1652630b	P G A M R Q L H W H P N A D E W . Q Y V L . . D G	20	G D . . V G Y V P K G Y G H A I R N	394]
1652726	P G A G P L P H I H H Y T N E W F W T P K . . G G	34	N Q . . I V Y G P K Y R V H G F A N	267
1652724	P G G G P M P H I H Y F I N E W F W T P E . . G G	34	K Q . . L I Y S P N H Y M H G F V N	268
1652717	P N A G P P P H F H Y A S D E W F F S T E . . P S	36	G E . . I I F S P R G M V H Y F R N	289
consensus	P G R H H E G		G P G H N	

FIG. 1. Sequence alignment of the central core of cupin sequences from *Synechocystis* PCC6803. Sequences are denoted according to the GenBank gi identifier. The number followed by a and b denote the first and second domain in this two-domain protein.

TABLE 1. ANALYSIS OF CLOSEST NEIGHBORS FOR EACH OF THE CUPIN SEQUENCES FROM *SYNECHOCYSTIS* PCC6803[a]

Sequence	*Closest neighbor*		*Length*	*Identity*	*Similarity*	*Gaps*	
gi	*gi*	*Species*	*AA*	*%*	*%*	*%*	*Function*
1652237	1657504	*E. coli*	47	34	55	–	Transcription regulator
1652486	1001180	*Synechocystis*	115	61	74	–	PMI
1001180	1652486	*Synechocystis*	102	63	78	–	PMI
1652906	1169942	*M. crystallinum*	95	27	50	9	GLP
1653078	347174	*Rhodococcus*	136	30	46	6	Cytochrome
1652749	1176281	*E. coli*	233	48	63	2	Pirin
1653678	141363	*S. enterica*	169	65	78	<1	Epimerase
1651977	1361427	*S. glaucescens*	179	42	55	–	Dehydrogenase
1652726	1652724	*Synechocystis*	233	48	68	–	?
1652724	1652726	*Synechocystis*	233	45	65	–	?
1652717	1652724	*Synechocystis*	132	26	47	6	?
1652630	1604990	*C. velutipes*	325	38	55	3	Oxalate decarboxylate

[a]Estimated by use of the gapped BLASTP program, showing the length of the region of greatest similarity, the percentages of identical and similar residues, and the percentages of gaps needed to give maximum similarity (TBLASTN was used for the two-domain GI|1652630 sequence, which has a DNA sequence as its neighbor).

of nuclear proteins, the pirins (Fig. 2) (Wendler et al., 1997). Additionally, gi|1653078 is very similar to the *Rhodococcus* sequence gi|347174, a putative C551 cytochrome, and gi|1657504 is similar to the *E. coli* transcriptional regulator gi|1657504. A lower degree of similarity, with gaps, is found between gi|1652906 and its nearest neighbor, the GLP (gi|1169942) from *Mesembryanthemum crystallinum.*

Although no function can be assigned yet to the remaining single-domain sequences, the two-domain 394-AA protein (gi|1652630) may be an oxalate decarboxylase, as predicted from its similarity (Fig. 3) to the 447-AA oxalate-degrading enzyme (Mehta and Datta, 1991) from the wood-rotting fungus *Collybia velutipes*. Its sequence (gi|1604990) has been published recently (Datta et al., 1996).

A detailed quantitative analysis of the cupin sequences shows a number of interesting features. If the single-domain proteins are considered first, in general terms the distance between motifs increases in line with the in-

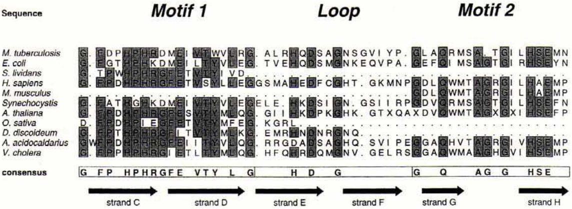

Sequence
Motif 1
Loop
Motif 2
M. tuberculosis G. FDPHPHRDMEIVTWVLRG. ALRHQDSAGNSGVIYP. GLAQRMSA.TGILHSEMN
E. coli G. FGTHPHKDMEILTYVLEG. TVEHQDSMGNKEQVPA. GEFQIMSAGTGIRHSEYN
S. lividans G. TPWHPHRGFETVTYIVD
H. sapiens G. FPDHPHRGFETVSYLLEGGSMAHEDFCGHT. GKMNPGDLQWMTAGRGILHAEMP
M. musculus GDLQWMTAGRGILHAEMP
Synechocystis G. FATHGHKDMEIVTYVLEGELE. HKDSIGN. GSIIRPGDVQRMSAGTGILHSEFN
A. thaliana G. FPDHPHRGFESVTYMLQG. GIIHKDPKGHK. GTXQAXDVQWMTAGXGXIHSEFP
O. sativa D. FPDHPIEGFETVTYMFEG. KGRL
D. discoideum G. FPTHPHRGFITVTYMLKG. EMRHNDNRGNQ.
A. acidocaldarius GWFPDHPHRGFEIITYVLQG. RRGDADSAGHQ. SVIPEGGAQHVTAGRGIVHSEMP
V. cholera G. FPPHPHRGIETLTYMLQG. HFQHRDQMGNV. GELRSGGAQWMAAGHGVIHSEMP
consensus G FP HPHRGFE VTY L G H D G G Q AG G HSE
strand C
strand D
strand E
strand F
strand G
strand H

FIG. 2. Sequence alignment of the central core of pirins from a range of microbes, plants, and animals. The details of the sequences used are as follows: *Mycobacterium tuberculosis* (gi|2213518), *Escherichia coli* (gi|1789847), *Streptomyces lividans* (translation of part of the -ve strand of the actinophage phi C31 attachment site gi|48953), *Homo sapiens* (gi|1907076), *Mus musculus* (EST gi|1282795), *Synechocystis* (gi|1652749), *Arabidopsis thaliana* (EST gi|950773), *Oryza sativa* (EST gi|1631547), *Dictyostelium discoideum* (-ve strand 302-3 of upstream sequence of *spiA* gene, gi|1177288), *Alicyclobacillus acidocaldarius* (manually edited from -ve strand upstream sequence of amylase gene gi|39300), *Vibrio cholerae* (sequence included in 538-nucleotide unfinished fragment gnl|TIGR|GVCCX37R).

```
                                                          Motif 1
I25120 111 FSFSKQRL--QTGGWARQQNEVVLPLATNLACTNMRLEAGAIRELHWHKN-AEWAYVLKG 167
           ++FSK  L    GG  +Q      P++  +A   M LE GAIRELHWH N AEWAYV++G
Synec.  56 YAFSKTPLVLYDGGTTKQVGTYNFPVSKGMAGVYMSLEPGAIRELHWHANAAEWAYVMEG 115

                                   Motif 2
I25120 168 STQISAVDNEGRNYISTVGPGDLWYFPPGIPHSLQATADDPEGSEFILVFDSGAFNDDGT 227
            T+I+    EG+  I+ V  G LWYFP G  HS++     P+ ++F+LVF+ G F++  T
Synec. 116 RTRITLTSPEGKVEIADVDKGGLWYFPRGWGHSIEGIG--PDTAKFLLVFNDGTFSEGAT 173


I25120 228 FLLTDWLSHVPMEVALKNFRAKNPAAWSHIPAQQLYIFPSEPPADNQPDPVSPQG---TV 284
           F +TDWLSH P+    +N      A  + +P +Q+YI  S  PA       +PQG    +
Synec. 174 FSVTDWLSHTPIAWVEENL-GWTAAQVAQLPKKQVYI-SSYGPASGPLASATPQGQTAKI 231
                                                             Motif 1..
I25120 285 PLPYSFNFSSVEPTQYSGGT-AKIADSTTFNISVAIAVAEVTVEPGALRELHWHPTEDEW 343
            +P++ N    +P    GG   ++A + F  S  +  A + +EPGA+R+LHWHP  DEW
Synec. 232 EVPHTHNLLGQQPLVSLGGNELRLASAKEFPGSFNMTGALIHLEPGAMRQLHWHPNADEW 291

           ..Motif 1                      Motif 2
I25120 344 TFFISGNARVTIFAAQSVASTFDYQGGDIAYVPASMGHYVENIGNTTLTYLEVFNTDRFA 403
            + + G   +T+FA++  AS    Q GD+ YVP   GH + N     L  + VFN    +
Synec. 292 QYVLDGEMDLTVFASEGKASVSRLQQGDVGYVPKGYGHAIRNSSQKPLDIVVVFNDGDYQ 351


I25120 404 DVSLSQWLALTPPSVVQAHLNLDDETLAEL 433
            + LS WLA  P SV+      +  E   +L
Synec. 352 SIDLSTWLASNPSSVLGNTFQISPELTKKL 381
```

FIG. 3. Protein sequences of the two-domain proteins from *Synechocystis* PCC6803 (gi|1652630) and *Collybia velutipes* (I25120), compared by gapped BLASTP. Score = 210 bits (558), expect = le-53, identities = 123/330 (37%), positives (+) = 179/330 (53%), gaps = 11/330 (3%). The shaded boxes denote motifs 1 and 2 within each domain.

crease in overall size of the protein (Fig. 1), from a minimum of 13 residues in the 129-AA gi|1652486 to 36 residues in the 289-AA gi|1652717. The one notable exception to this trend is the 232-AA gi|1652749, which has a 16-residue intermotif loop. As reported, the two-domain sequence is related to a fungal decarboxylase. Both have 20 residues between motifs, a spacing characteristic of many of the GLPs from higher plants.

DISCUSSION

Before the present study, only four of the *Synechocystis* sequences had been identified as cupins (Dunwell and Gane, 1998). Three of these were single-domain proteins, namely, the PMI gi|1001180 and its two related sequences, gi|1652906 and gi|1653708. The other was the two-domain protein gi|1652630. The total number in this paralogous family is now increased to 12 (Fig. 1 and Table 1), although not all of these proteins fulfil the arbitrary definition used in the recent analysis of the *E. coli* genome (Blattner et al., 1997). These authors used the term to include all those ORFs that share at least 30% sequence identity over more than 60% of their lengths. However, it is probably not appropriate to apply this strict criterion to the present group of 12 sequences, which vary in length from 104 to 394 AAs and in which the two most conserved motifs are separated by a nonconserved region of variable length.

Four of the proteins identified in the present study have a possible functional connection, in that they are concerned with cell wall synthesis. The best known are the PMIs (EC 5.3.1.8), enzymes that catalyze the interconversion of mannose-6-phosphate and fructose-6-phosphate. Although they are considered to be zinc-containing metalloproteins, an Fe(III)-hydroxyphenylalanine site has been identified recently in the PMI from *Candida albicans* when expressed in *E. coli* (Proudfoot et al., 1996; Smith et al., 1997). On the basis of sequence comparison, PMIs have been divided into three classes (Proudfoot et al., 1994), within which the class II enzymes (those described here) are involved in a variety of pathways, including capsular polysaccharide biosynthesis and D-mannose metabolism. Interestingly, in *Synechocystis* there seems to be no example of the bifunctional GDP-mannose pyrophosphorylase/PMI enzyme (the PMI domain of about 130 AAs is located at the C-terminus of the protein) found in some Archaea (e.g., the 448-AA gi|2649495 from *Archaeoglobus fulgidus*) and many eubacteria (e.g., the 428-AA gi|1230580 from *Vibrio cholerae* and the 470-AA gi|2313118 from *Helicobacter pylori*) and thought to be involved, for example, in the polymerization of alginate, a viscous mucoid exopolysaccharide. Instead, the two functions in *Synechocystis*, as in *Methanococcus jannaschii* (J.M. Dunwell, unpublished observations), are carried out by two separate enzymes, the phosphorylase function by the 367-AA gi|1653346 and the isomerase function by the two PMIs (approximately 128 AA), gi|1001180 and gi|1652486, identified previously. Other bacteria, such as *E. coli*, contain a family of related enzymes, including PMIs (e.g., the 152-AA gi|147164), as well as several bifunctional enzymes (e.g., the 471-AA gi|585853, the 478-AA gi|1155018, and the 483-AA gi|1584629).

Enzymes, such as the dTDP-4-dehydrorhamnose 3,5-epimerase (gi|1653678) identified here, are involved in the synthesis of bacterial exopolysaccharides. These enzymes include the ExpA8 protein recently shown to be involved in the synthesis of galactoglucan (exopolysaccharide II) in *Rhizobium melioti* (Becker et al., 1997) and the TDP-deoxyglucose epimerase (L33181) that is part of the biosynthetic pathway for ascarylose, a lipopolysaccharide component in *Yersinia pseudotuberculosis* (Thorson et al., 1994). The related dTDP-6-deoxy-L-mannose-dehydrogenase (gi|1651977), identified in this study as a cupin, is also involved in cell wall synthesis.

In contrast to the presumed function in cell wall synthesis assigned to the previous four sequences considered, the identity of sequence gi|1652749 seems to be as the nuclear protein pirin, one of a highly conserved group of proteins (Fig. 2) thought to interact with the nuclear factor I/CCAAT box transcription factor (NFI/CTF1) (Wendler et al., 1997). Of the other sequences, gi|1652237 has the transcriptional regulator gi|1657504 from *E. coli* as its closest neighbor, and gi|1653078 is similar to gi|347174, a cytochrome C551 gene from *Rhodococcus*.

Unfortunately, no function can be assigned to any of the other single-domain cupins listed in Figure 1, although it is hoped that as this superfamily is analyzed in more detail, the subgroups will be identified systematically by their homology to proteins of known function—a process that enabled the oxalate oxidase identity of the wheat gf 2.8 germin to be established (Lane et al., 1993).

The detailed analysis summarized in Figure 1 confirmed that all the cupin sequences found in *Synechosystis* contain the characteristic two-motif structure, with the intermotif distance varying in length from 13 (a uniquely short spacing for this superfamily) to 34 residues. Despite the range of spacing found in this study, it is interesting to note that none of the single-domain cupins in *Synechocystis* has an intermotif distance of 20 residues, the spacing found in the two-domain sequence gi|1652630. There are a number of possible explanations for this peculiarity. First, it may be that the duplication occurred in a protein with a 16-residue spacing, followed by the insertion of 4 residues in each domain; this seems unlikely. Alternatively, the 20-residue progenitor may have been lost through natural selection or may simply not have been identified in this study. Again it seems unlikely that any other cupin sequences in this genome remain undiscovered. In a further attempt to find close relatives to the two-domain 20-residue protein from other organisms, the 57-AA sequence spanning the two motifs of domain 1 was used as a probe in a BLASTP search. This revealed an *Arabidopsis* GLP (U75207) as the closest neighbor, with the closest nonplant sequence being the hypothetical protein gi|2128971 from the archaeon *M. jannaschii* (37% identical, 48% similar over a distance of 51 AAs). However, this 125-AA protein has only a 16-residue gap. Similarly, use of an equivalent sequence from domain 2 revealed the 113-AA sequence gi|1881251 from *Bacillus subtilis* as the nearest neighbor (37% identical, 62% similar over 51 residues). This sequence also has a 16-AA gap. Therefore, to date, there is no known example of a single-domain, 20-AA gap protein from a prokaryote. This apparent lack of any progenitor 20 protein, allied to the multitude of prokaryotic two-domain proteins with the 20 spacing, suggests that there was only one 20 protocupin, which underwent a duplication to produce gi|1652630 and its equivalents in other species and left no extant progenitor (or at least none identified to date).

Comparison of the alignment of the two-domain sequences from *Synechocystis* and *C. velutipes* (Fig. 3) confirms that they are probably derived from the same progenitor and that the former sequence may be the direct progenitor of the latter. This view is reinforced by analysis not only of the conserved motif regions but also of the intervening intermotif regions. Specifically, there are several identical residues in the two

20-AA intermotif regions of the first domain of each protein (consensus XTXIXXXXXEGXXXIXXVXX) and a different set of identical residues (consensus XXXXTXFAXXXXASXXXXQX) in the intermotif regions of the second domain. This suggests strongly that there was divergence of the two domains of a precursor protein after duplication of a 20-residue protocupin, followed by further minor divergence during the postduplication phase, leading eventually to the present day sequences.

In terms of both their number and their range of size and the presence of a two-domain sequence, the evidence presented suggests that the spectrum of cupins found in *Synechocystis* is closer to that of higher plants rather than to that found in more primitive bacteria (to date, only two cupins have been identified in *M. jannaschii*) (Dunwell, 1997). Presumably, there was a rapid expansion of cupin diversity during the evolution of *Synechocystis*. This conclusion may be related to the observation that the genome of this organism contains 99 ORFs with similarity to transposases (Kaneko et al., 1996). It was suggested that this high frequency could be linked to frequent rearrangement of the genome during and after establishment of this species. More recently, Cassier-Chauvat et al. (1997) characterized three specific insertion sequences from *Synechocystis* and suggested, on the basis of homology, that they were spread through horizontal transfer between evolutionarily distinct organisms.

Identification of the prokaryotic two-domain sequence in the present study also complicates the conclusion reached by Bäumlein et al. (1995) that "the extant spherulins and germins might represent a stage of seed globulin gene evolution *before* [my emphasis] the domain duplication event had occurred."

If the linear structure of these 12 *Synechocystis* cupin sequences is considered, it can be seen that there is an overall increase in spacing between the motifs, coincident with the increase in length of the proteins (Fig. 1). If it is assumed that the progenitor of these proteins was the smallest and most simple of the protocupins, as the overall sequence grew in length by addition of residues at each end, this must have been accompanied by insertion of residues into the variable region, namely, the E/F loop at the end of the β-barrel (Gane et al., 1997). This gradual increase in complexity led eventually to addition of α-helical regions at each end of the protein, duplication into two long $\alpha + \beta$ domains, and finally to assembly of the protein subunits to give the trimeric quaternary structure found in the abundant storage proteins of land plants (Lawrence et al., 1994) (Table 2). In addition, by the present stage of the evolutionary process, two of the three conserved histidines have been lost in most storage proteins (all three in some examples), and as these residues are implicated as being catalytically active (Gane et al., 1998), it is likely that these multimeric proteins no longer have any enzyme activity (none is known). In this regard, it is also not known if the *Synechocystis* two-domain protein or any of the more primitive two-domain storage proteins show oxalate decarboxylase activity (cf. I25120).

Duplication of domains, followed by their subsequent divergence during evolution, is known to occur in other proteins, such as the zinc metalloenzyme glyoxylase 1, which catalyzes the glutathione-dependent inactivation of toxic methylglyoxyl (Cameron et al., 1997). In this example, sequence alignment showed that 13 of 74 residues were identical in domain 1 and 13 of 59 in domain 2. Such relatively low levels of identity occur presumably because of the small number of residues required to provide the conserved structural elements in such proteins. Undoubtedly, as analysis techniques become more sophisticated, more cases of

TABLE 2. CHARACTERISTICS OF A SAMPLE OF CUPIN PROTEINS FROM THE EVOLUTIONARY SEQUENCE FROM BACTERIA TO HIGHER PLANTS[a]

Name	*Species*	*Length AA*	*Histidines*	*Loop AA*	*Domains*	*Subunits*
PMI	*Synechocystis*	128	3	13	1	1
GLP	*A. thaliana*	200	3	20	1	1
Oxalate oxidase	*T. aestivum*	201	3	23	1	?5
Oxalate decarboxylate	*C. velutipes*	447	3	20	2	1
Vicilin	*P. vulgaris*	397	1	27	2	3

[a] The proteins are classified according to their name, species of origin, total length, number of conserved histidine residues in the two motifs, length of the intermotif loop, number of domains, and number of subunits in the mature protein.

this type of evolutionary process will be revealed, and the number of basic underlying 3D structures will be found to be restricted (Godzik, 1997).

ACKNOWLEDGMENTS

I would like to acknowledge the financial support of Zeneca plc and the Biotechnology and Biological Sciences Research Council and the expertise of Paul Gane for preparation of the figures.

REFERENCES

ALTSCHUL, S.F., MADDEN, T.L., SCHAFFER, A.A., ZHANG, J., ZHANG, Z., MILLER, W., et al. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programmes. Nucleic Acids Res **25,** 3389–3402.

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

BECKER, A., RUBERG, S., KUSTER, H., ROXLAU, A.A., KELLER, M., IVASHINA, T., et al. (1997). The 32-kilobase *exp* gene cluster of *Rhizobium melioti* directing the biosynthesis of galactoglucan: Genetic organization and properties of the encoded gene products. J Bacteriol **179,** 1375–1384.

BLATTNER, F.R., PLUNKETT, G., BLOCH, C.A., PERNA, N.T., BURLAND, V., RILEY, M., et al. (1997). The complete genome sequence of *Escherichia coli* K-12. Science **277,** 1453–1462.

CAMERON, A.D., OLIN, B., RIDDERSTRÖM, M., MANNERVIK, B., and JONES, T.A. (1997). Crystal structure of human glyoxylase 1—Evidence for gene duplication and 3D domain swapping. EMBO J **16,** 3386–3395.

CASSIER-CHAUVAT, C., PONCELET, M., and CHAUVAT, F. (1997). Three insertion sequences from the cyanobacterium *Synechocystis* PCC6803 support the occurrence of horizontal DNA transfer among bacteria. Gene **195,** 257–266.

DATTA, A., MEHTA, A., and NATARAJAN, K. (1996). Oxalate Decarboxylase. US Patent 5547870.

DUNWELL, J.M. (1997). Cupins: A new superfamily of functionally diverse proteins that include germins and plant storage proteins. Biotech Genet Eng Rev **15,** 1–32.

DUNWELL, J.M., and GANE, P.J. (1997). Microbial relatives of seed storage proteins: Conservation of motifs in a functionally diverse superfamily of enzymes. J Mol Evol **45,** 147–154.

GANE, P.J., DUNWELL, J.M., and WARWICKER, J. (1997). Modeling based on the structure of vicilins predicts a histidine cluster in the active site of oxalate oxidase. J Mol Evol **45,** 488–493.

GODZIK, A. (1997). Counting and classifying possible protein folds. Tibtech **15,** 147–151.

HENIKOFF, S., and HENIKOFF, J.G. (1991). Automated assembly of protein blocks for database searching. Nucleic Acids Res **19,** 6565–6572.

KANEKO T., SATO, S., KOTANI, H., TANAKA, A., ASAMIZU, E., NAKAMURA, Y., 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.

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

LAWRENCE, M.C., IZARD, T., BEUCHAT, M., BLAGROVE, R.J., and COLMAN, P.M. (1994). Structure of phaseolin at 2.2 A resolution: Implications for a common vicilin/legumin structure and the genetic engineering of seed storage proteins. J Mol Biol **238,** 748–776.

LEHEL, C., LOS, D., WADA, H., GYORGYEI, J., HORVATH, I., KOVACS, E., et al. (1993). A second *groEL*-like gene, organized in a *groESL* operon is present in the genome of *Synechocystis* sp. PCC 6803. J Biol Chem **268,** 1799–1804.

MEHTA, A., and DATTA, A. (1991). Oxalate decarboxylase from *Collybia velutipes*. Purification, characterization and cloning. J Biol Chem **266,** 23548–23553.

PROUDFOOT, A.E., GOFFIN, L., PAYTON, M.A., WELLS, T.N., and BERNARD, A.R. (1996). In vivo and in vitro folding of a recombinant metalloenzyme, phosphomannose isomerase. Biochem J **318,** 437–442.

PROUDFOOT, A.E.I., TURCATTI, G., WELLS, T.N.C., PAYTON, M.A., and SMITH, D.J. (1994). Purification, cDNA cloning and heterologous expression of human phosphomannose isomerase. Eur J Biochem **219,** 415–423.

SMITH, J.J., THOMSON, A.J., PROUDFOOT, A.E., and WELLS, T.N. (1997). Identification of an Fe(III)-dihydroxyphenylalanine site in recombinant phosphomannose isomerase from *Candida albicans*. Eur J Biochem **244,** 325–333.

THORSON, J.S., LO, S.F., PLOUX, O., HE, X., and LIU, H.W. (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.

WENDLER, W.M.F., KREMMER, E., FOERSTER, R., and WINNACKER, E.L. (1997). Identification of pirin, a novel highly conserved nuclear protein. J Biol Chem **272,** 8482–8489.

Address reprint requests to:
Jim M. Dunwell
Department of Agricultural Botany
School of Plant Sciences
The University of Reading
Whiteknights PO Box 221
Reading RG6 6AS
UK