Available online at www.sciencedirect.com
SCIENCE DIRECT®
PHYTOCHEMISTRY
Phytochemistry 65 (2004) 7–17
www.elsevier.com/locate/phytochem

Molecules of Interest

Cupins: the most functionally diverse protein superfamily?

Jim M. Dunwell*, Alan Purvis, Sawsan Khuri

The BioCentre, School of Plant Sciences, The University of Reading, Whiteknights, Reading RG6 6AS, UK

Accepted 13 August 2003

Abstract

The cupin superfamily of proteins, named on the basis of a conserved β-barrel fold ('*cupa*' is the Latin term for a small barrel), was originally discovered using a conserved motif found within germin and germin-like proteins from higher plants. Previous analysis of cupins had identified some 18 different functional classes that range from single-domain bacterial enzymes such as isomerases and epimerases involved in the modification of cell wall carbohydrates, through to two-domain bicupins such as the desiccation-tolerant seed storage globulins, and multidomain transcription factors including one linked to the nodulation response in legumes. Recent advances in comparative genomics, and the resolution of many more 3-D structures have now revealed that the largest subset of the cupin superfamily is the 2-oxyglutarate-Fe^{2+} dependent dioxygenases. The substrates for this subclass of enzyme are many and varied and in total amount to probably 50–100 different biochemical reactions, including several involved in plant growth and development. Although the majority of enzymatic cupins contain iron as an active site metal, other members contain either copper, zinc, cobalt, nickel or manganese ions as a cofactor, with each cofactor allowing a different type of chemistry to occur within the conserved tertiary structure. This review discusses the range of structures and functions found in this most diverse of superfamilies.

Keywords: Germin; Dioxygenase; Auxin-binding protein; Pirin; Gibberellin; Legumin; Vicilin; Metalloenzymes

1. Introduction

Identification of the cupin superfamily was originally based on the realisation that the wheat protein germin, an unusual thermostable protein produced during the early phase of germination in wheat embryos, shared a nine amino acid sequence (HI/THPRATEI) with a stress-related protein, a spherulin, produced during starvation of the slime mould *Physarum polycephalum*. This similarity was then extended to a group of germin-like proteins (GLPs) in dicotyledonous plants, and subsequent comparative sequence analysis noted a much weaker, though consistent, level of similarity to the globulin storage proteins such as the desiccation-tolerant vicilins and legumins from plant seeds and spores. Knowledge of the 3D structure of these seed proteins enabled a more detailed structure-based alignment to be produced and this revealed a larger grouping of proteins, including many microbial examples (Dunwell and Gane, 1998); collectively they were given the annotation of cupins on the basis of their β-barrel shape ('cupa' is the Latin term for small barrel) (Dunwell, 1998). The characteristic cupin domain comprises two conserved motifs, each corresponding to two β-strands, separated by a less conserved region composed of another two β-strands with an intervening variable loop. The total size of the intermotif region varies from a minimum of 11 AAs in some microbial enzymes, to ca. 50 AAs in the non-enzymatic seed storage proteins, and to >100 AAs in certain eukaryotic transcription factors and dioxygenases (Dunwell et al., 2000, 2001). For Motif 1 the characteristic conserved sequence was originally designated as $G(X)_5HXH(X)_{3,4}E(X)_6G$ and $G(X)_5PXG(X)_2H(X)_3N$ for Motif 2 (Fig. 1); these two motifs were defined before confirmation (Woo et al., 2000) of the previous prediction (Gane et al., 1998) that the two His residues and the Glu residue in Motif 1, together with the His residue in Motif 2, acted as ligands for the binding of the active site metal manganese ion in the archetypal cupin, germin (Fig. 2a)

* Corresponding author. Tel. +44-118-378-6313; fax: +44-118-378-8160.

0031-9422/$ - see front matter © 2003 Elsevier Ltd. All rights reserved.
doi:10.1016/j.phytochem.2003.08.016

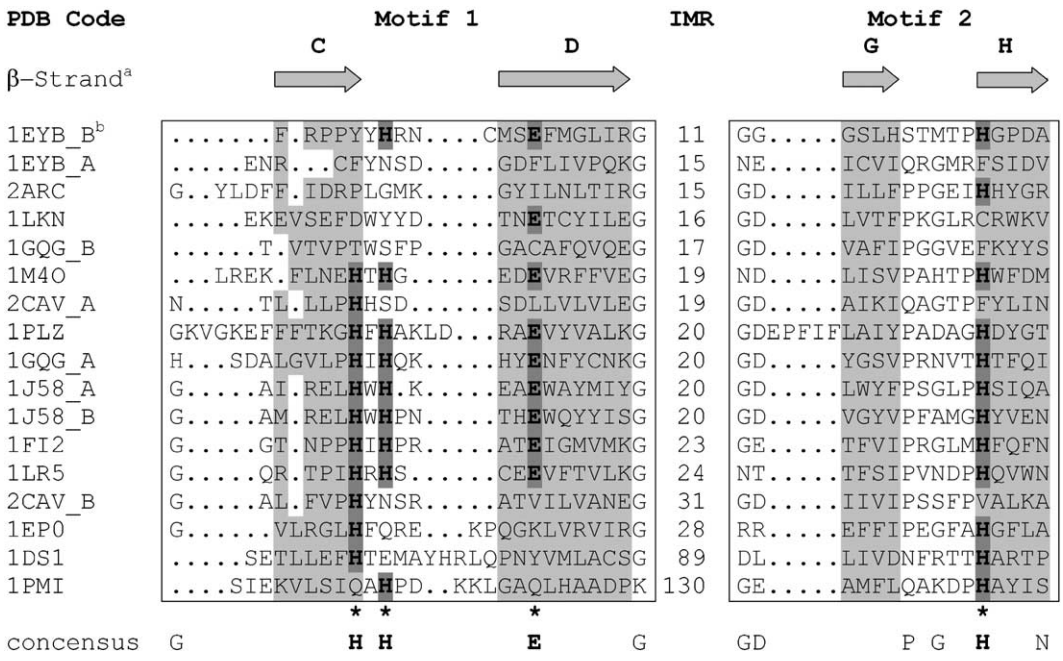

Fig. 1. Multiple alignment of representative cupin proteins (structures as shown in Fig. 2b), showing the two diagnostic conserved domains, the four β-strands (shaded), the number of AAs in the inter-motif region (IMR), the concensus sequence, and the amino acids (in bold) that act as ligands for the active site metal cofactor. [a]The four β-strands, C, D, G, H are labelled using the convention based on the structure of phaseolin (Lawrence et al., 1994). [b]The A or B suffix following the four-symbol PDB code designates either the N- or C-terminal domain, respectively, of two-domain bicupins.

(see below). It is now clear that the primary sequence of two motifs is much less conserved than first suggested. For example, one of the two His residues in Motif 1 is substituted by Gln or Asp/Glu in some isomerases or dioxygenases (Fig. 1).

The various enzymatic and non-enzymatic functions will be described below in an overall survey of the cupin superfamily. It has been estimated previously that there are minimum of 18 different functional subclasses (Dunwell et al., 2001), but this figure is now known to be much higher. However, experimental biochemical evidence can only be provided for a minority of these classes. It should be emphasised that much of the recent information on this superfamily has been derived from the rapidly increasing number of cupin structures (Fig. 2b) that have been resolved by X-ray crystallography or NMR techniques (Table 1), and by the subsequent use of structure-based alignments. Data from these studies have confirmed many of the early predictions about possible members of the superfamily, but also have extended membership to proteins with very limited primary sequence similarity and not included in the original reviews.

The different major cupin subgroups will be considered in turn below, based upon whether the proteins comprise a single cupin domain, or whether they have a duplicated (bicupin) or multicupin (>2 cupin domains) structure. In a review of this length, only a minimum of information can be provided about each type; emphasis will be given to recent publications and to proteins known to be present in plants or in microbes associated with plants. The reader is referred to previous reviews (Dunwell et al., 2000, 2001) for more detail.

2. Monocupins

These are proteins that include a single cupin domain, either at the centre of a simple protein (ca. 80–200 AAs in length) or as one of several domains in a complex protein (200–ca. 1000 AAs). They mostly comprise enzymes, but microbial non-enzymatic transcription factors (eg AraC), in which a sugar-binding effector domain is linked to a DNA-binding domain, are also well known.

2.1. Dioxygenases

Of all the various types of cupin, the most abundant generic class comprises dioxygenases, enzymes that catalyze a variety of reactions typically involving the oxidation of an organic substrate using a dioxygen molecule. Most of the varied subtypes contain iron as the active site metal cofactor but other forms are known, including examples with copper or nickel (Table 1) (see below).

2.1.1. 2-Oxoglutarate iron dependent dioxygenases

The predominant subclass, with the most varied range of substrates, are the 2-oxoglutarate (2OG) and

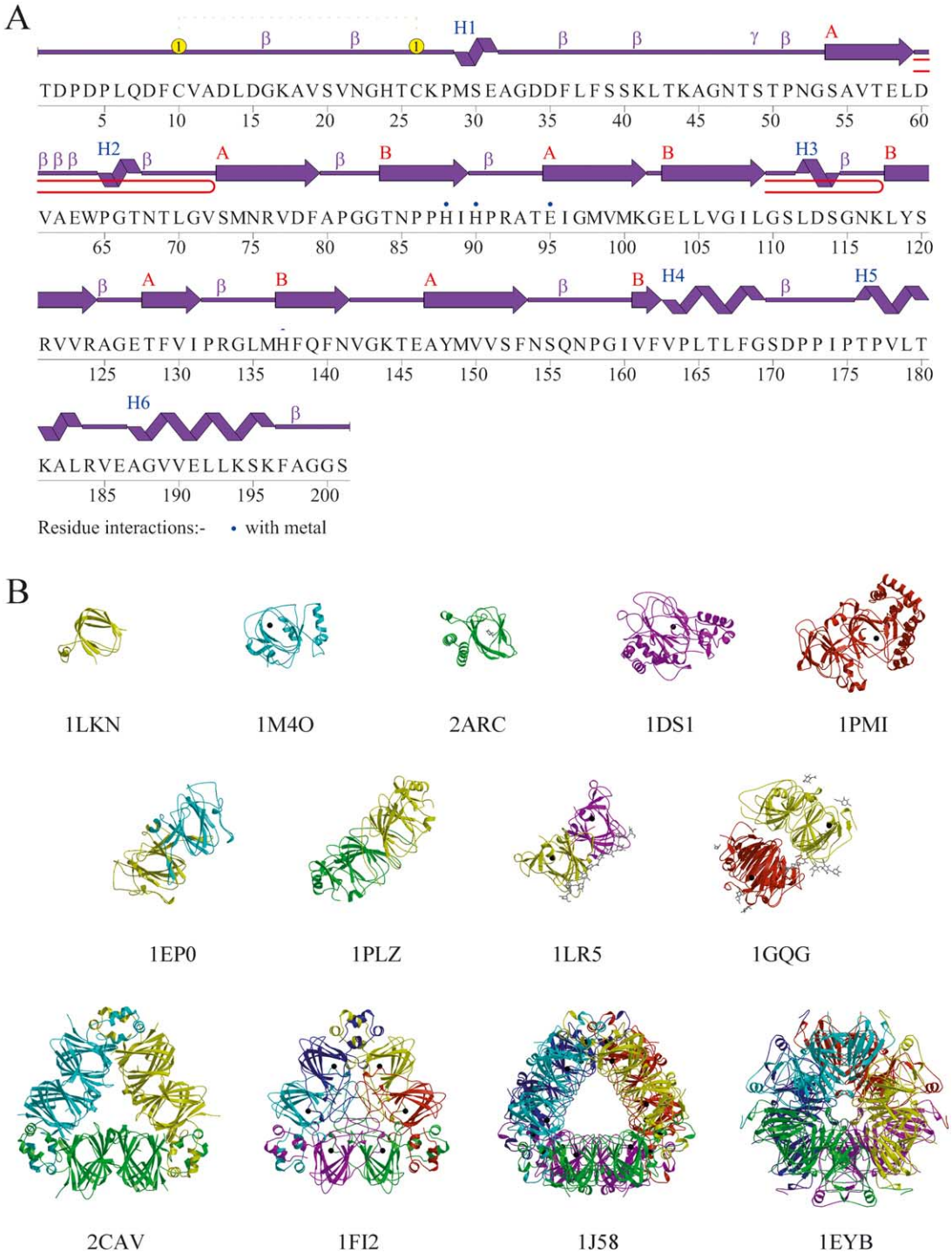

Fig. 2. (A). The primary and secondary structure of the archaetypal cupin, germin (PDB code 1FI2), from the PDBsum database (http://www.biochem.ucl.ac.uk/bsm/pdbsum/index.html) showing the position of the two sheets (A, B in red), six helices (H1-6) and 10 strands, the two β-hairpins (in red), 17 β-turns (purple), one γ-turn, four residues (H88, H90, E95, H137) that act as ligands for the active site Mn ion, and the position of the disulphide bond between residues 10 and 26 (yellow) within the N-terminal extension. (B) Representative selection of cupin structures showing the PDB code (http://www.rcsb.org/pdb/index.html), and the multimeric composition. Further details of each protein are given in Table 1.

Fe^{2+}-dependent oxygenases, a class of enzymes that are widespread in eukaryotes, bacteria, and certain plant RNA viruses, though notably not in the Archaea (Aravind and Koonin, 2001). Collectively, these enzymes have roles in acquisition of nutrients and synthesis of antibiotics in bacteria and secondary products in plants, in repair of alkylated nucleic acids in both prokaryotes and eukaryotes, and in the synthesis of collagen and sensing of oxygen in mammals; some of these functions will be defined below.

Table 1
Summary of cupin structures resolved using X-ray crystallography or NMR techniques. A representative sample of these structures is illustrated in Fig. 1b

Protein	Species	Kingdom[a]	PDB code[b]	Metal ion	Inter-motif length (aa)
Acireductone dioxygenase	*Klebsiella pneumoniae*	B	1M4O[c]	Ni	19
Anthocyanidin synthase	*Arabidopsis thaliana*	E	1GP4	Fe	23
Auxin binding protein	*Zea mays*	E	1LR5	Zn	24
Canavalin	*Canavalia ensiformis*	E	2CAV	None	19/31[d]
Clavaminate synthase	*Streptomyces clavuligerus*	B	1DS1	Fe	89
Deacetoxycephal. synthase	*Streptomyces clavuligerus*	B	1DCS	Fe	34
dTDP dehydrorhamnose epim.	*Salmonella typhimurium*	B	1DZR	None	28
	Methanobact. thermoautotr.	A	1EP0	None	28
GAB protein	*Escherichia coli*	B	1JR7	Fe	105
Germin	*Hordeum vulgare*	E	1FI2	Mn	23
Homogentisate dioxygenase	*Homo sapiens*	E	1EYB	Fe	15/11[d]
Isopenicillin N synthase	*Emericella nidulans*	E	1BK0	Fe	34
Oxalate decarboxylase	*Bacillus subtilis*	B	1J58	Mn	20/20[d]
Phaseolin	*Phaseolus vulgaris*	E	2PHL	None	25/27[d]
Phosphoglucose isomerase	*Pyrococcus furiosus*	A	1PLZ	Fe?	20
Phosphomannose isomerase	*Candida albicans*	E	1PMI	Zn	130
Proglycinin	*Glycine max*	E	1FXZ	None	42/21[d]
Proline-3-hydroxylase	*Streptomyces*	B	1E5S	Fe	27
Quercetin dioxygenase	*Aspergillus japonicus*	E	1GQG	Cu	20/17[d]
Taurine-α-ketoglut. diox.	*E. coli*	B	1OS7	Fe	126
Transcription factor	*E. coli*	B	2ARC	None	15
Unknown	*Thermotoga maritima*	B	1LKN[b]	?	16

[a] A, archaeon; B, bacterium; E, eukaryote.
[b] Only one Protein Data Bank (PDB) accession code is given for each protein, although there may be several structures deposited in this database.
[c] These structures were resolved using NMR techniques; the remainder were resolved using X-ray crystallography.
[d] These two values relate to the N- and C-terminal cupin domains, respectively, in a bicupin protein.

Within fungi and plants, the largest group of this type of dioxygenase (>50 in the Arabidopsis genome) is that represented by domain 6999 from the conserved domain database (CDD), equivalent to pfam03171. The concensus primary sequence for this class is: RAESLLVNRYPELGLGLGP**H**T**D**DEGQRILTILL-QDGGGGLQFPKDDGWIDVPPEPGALLVNFGD-LLEMLNGRYKSVL**H**RVLPVTGGTRISLAFFLRP (the three iron-binding ligands in Motifs 1 and 2 are shown in bold, see Fig. 1).

Although most of these proteins do not have any confirmed biochemical function, some have been the subject of detailed biochemical study. Such plant enzymes include: 1-aminocyclopropane-1-carboxylate oxidase (EC 1.4.3.-); anthocyanidin synthase, also known as leucoanthocyanidin dioxygenase (EC 1.14.11.-) (PDB code:1GP4, Table 1); flavanone 3-hydroxylase; naringenin 3-dioxygenase (EC 1.14.11.9); desacetoxyvindoline 4-hydroxylase; gibberellin 20-oxidase (EC 1.14.11.-); gibberellin 7-oxidase (EC

1.14.11.-); gibberellin 2 β-hydroxylase; gibberellin 3 β-hydroxylase (EC 1.14.99.-); gibberellin 3 β-dioxygenase (EC 1.14.11.15); hyoscyamine 6 β-hydroxylase; hyoscyamine (6S)-dioxygenase (EC 1.14.11.11). In the area of plant breeding, the enzyme of this class with possibly the greatest significance is that responsible for the major semi-dwarfing allele (*sd-1*) in rice; this allele was used to generate the well-known varieties IR 8 and IR36, and is still extensively used in modern rice cultivars. Recent analysis of several semi-dwarf rice lines by two independent groups (Spielmeyer et al., 2002; Monna et al., 2002) have shown that the *sd-1* phenotype is caused by a deficiency of bioactive gibberellins and have identified various related mutations in the GA 20-oxidase gene, the product of which converts GA_{53} to GA_{20}. Specifically, a deletion of 280 bp within the coding region of the *sd-1* (*Os20ox2*) gene was predicted to encode a nonfunctional protein in an indica type semi-dwarf (Doongara), whereas a substitution at amino acid position 266 (Leu [CTC] to Phe [TTC] within the cupin inter-motif region) could explain loss of function of the same gene in a japonica semi-dwarf (Calrose76). In an associated study, quantification of GAs in elongating stems by GC-MS showed that the initial substrate of GA 20-oxidase activity (GA_{53}) accumulated, whereas the content of the major product (GA_{20}) and of bioactive GA_1 was lower in semi-dwarf compared with tall lines.

The only one of these plant enzymes to have a resolved tertiary structure is anthocyanidin synthase (ANS), the enzyme that catalyzes the penultimate step in the biosynthesis of the anthocyanin class of flavonoids. Together with in vitro studies, the crystal structures suggest a mechanism for ANS-catalyzed anthocyanidin formation from the natural leucoanthocyanidin substrates involving stereoselective C-3 hydroxylation (Wilmouth et al., 2002). The structure of ANS also provides a template of this ubiquitous family of plant nonhaem oxygenases for future engineering and inhibition studies.

One interesting plant specific enzyme of this type is that associated with the synthesis of benzoxazinoids, secondary metabolites of grasses that function as natural pesticides and form an important part of the defense against insects and microbial pathogens. Although various steps in the biosynthesis of one such well-known compound, the cyclic hydroxamic acid, 2,4-dihydroxy-7-methoxy-1,4-benzoxazin-3-one (DIMBOA) have been elucidated, the mechanism of the introduction of OCH_3-group at the C-7 position was unknown. Recently, however, inhibitor experiments in *Triticum aestivum* and *Zea mays* suggest that the enzyme responsible for this hydroxylation reaction at C-7 is a 2OG-dependent dioxygenase (Frey et al., 2003). Cloning and reverse genetics analysis also identified the *Bx6* gene that encodes this enzyme.

Amongst the most important group of these enzymes is that involved with the protection of nucleic acids against environmental damage. Specifically, the identification of AlkB, a bacterial protein known to be involved in cellular recovery from alkylation damage, as a member of the 2OG-Fe^{2+} oxygenase superfamily suggested that this protein catalyzes oxidative detoxification of alkylated bases, a suggestion supported by the identification of more distant homologs detected in eukaryotes and in plant RNA viruses (Aravind and Koonin, 2001). This conclusion lead to the hypothesis that these proteins might also be involved in RNA demethylation. Within the last 2 years, several experimental studies have confirmed this hypotheses. For example, Falnes et al. (2002) showed that AlkB from *Escherichia coli* releases replication blocks in alkylated DNA by a mechanism involving oxidative demethylation of 1-methyladenine residues. This mechanism represents a new pathway for DNA repair and is the third type of DNA damage reversal mechanism so far discovered to date. Similarly, Trewick et al. (2002) showed that purified AlkB repairs the cytotoxic lesions 1-methyladenine and 3-methylcytosine in single- and double-stranded DNA in a reaction that is dependent on oxygen, alpha-ketoglutarate and Fe^{2+}. The AlkB enzyme couples oxidative decarboxylation of alpha-ketoglutarate to the hydroxylation of these methylated bases in DNA, resulting in direct reversion to the unmodified base and the release of formaldehyde. In related study, it was recently shown that two human AlkB homologues, hABH2 and hABH3, are oxidative DNA demethylases and that AlkB and hABH3, but not hABH2, also repair RNA (Aas et al., 2003). These results collectively illustrate the biological relevance of this repair activity (Begley and Samson, 2003) and establish RNA repair as a potentially important defence mechanism in all living cells. Although there is no functional evidence from plants, many AlkB-type sequences are known, with the most closely related protein from Arabidopsis being gi|15221095.

In this context, the identification of an AlkB-type dioxygenase domain within a group (carla-, tricho- and potex families) of plant ssRNA positive–strand viruses (e.g. gi|23821049 from the Aconitum latent virus) is an intriguing discovery. First, this is the only example to date of a cupin protein encoded by any viral genome, and secondly it appears to be a good example of horizontal gene transfer of a plant defence gene to a plant pathogen. Presumably, this is part of an evolutionary strategy of the virus to overcome a plant-induced methylation of an invasive viral RNA.

Another major, and ubiquitous subgroup of this type is the prolyl and asparagyl hydroxylases, enzymes with several most important roles in cellular regulation. For example, in the sensing of endogenous oxygen tension in animals, activity of the hypoxia-inducible factor (HIF)

complex is controlled by oxygen-dependent hydroxylation of prolyl and asparaginyl residues. Specifically, hydroxylation of these residues by 2OG-dependent oxygenases mediates ubiquitinylation and proteasomal destruction of HIF-alpha (Hewitson et al., 2002). There are several plant homologues of these enzymes (e.g. gi|30685945 from *Arabidopsis*) and this suggests their possible involvement in responses to hypoxia (e.g. waterlogging).

A related, medically important example is provided by phytanoyl-CoA alpha-hydroxylase (EC 1.14.11.18), the mammalian peroxisomal enzyme responsible for the α-oxidation of phytanic acid (3,7,10,14-tetramethylhexadecanoic acid). This branched-chain fatty acid is commonly found in the human diet and may be derived from chlorophyll in plant extracts. Lack of the oxidative enzyme, leading to accumulation of phytanic acid, is found in genetic disorders such as Refsum's disease (hereditary motor sensory neuropathy type IV, heredopathia atactica polyneuritiformis), an autosomal recessive condition leading to neurological distress, deterioration of vision, deafness, loss of coordination and eventual death.

Another putative dioxygenase sequence is the jmjC domain identified in a large family of eukaryotic and bacterial transcription factors (Clissold and Ponting, 2001). These include: *Vicia faba* early nodulation binding protein (ENBP1) and its relatives, human hairless (mutated in individuals with *alopecia universalis*), retinoblastoma-binding protein 2 and several chromatin-associated proteins.

2.1.2. Acireductone dioxygenases

Amongst the most interesting of the cupin structures solved recently (Pochapsky et al., 2002) is that of the *Klebsiella pneumoniae* acireductone dioxygenase (ARD), the first determined for this new family of metalloenzymes. ARD represents a branch point in the methionine salvage pathway leading from methylthioadenosine to methionine and is most unusual in that it has been shown to catalyze different reactions depending on the type of metal ion bound in the active site; with the 5-methylthio-D-ribose moiety of *S*-methylthioadenosine as a substrate, addition of Ni^{2+} or Co^{2+} to the apo-protein yields E-2 activity (a 1,3-oxygenlytic reaction), whereas E-2′ activity (a 1,2-oxygenlytic reaction) is obtained when Fe^{2+} is added. The solution structure of nickel-containing version was recently determined using NMR methods (Fig. 2b). X-ray absorption spectroscopy, assignment of hyperfine shifted NMR resonances and conserved domain homology were used to model the metal-binding site because of the paramagnetism of the bound Ni^{2+}. Plant versions of this ubiquitous enzyme include the so-called submergence induced protein from rice (e.g. gi|2952338).

2.2. Germin and GLPs

The similarity of a short motif in germin to a sequence from a slime mould spherulin was the key to the discovery of the cupin superfamily (Dunwell, 1998). Since that time the number of germin-like proteins (GLPs) and their significance has increased rapidly; with the exception of the 2OG-Fe^{2+} dioxygenases (see above) they now represent the largest gene family of any cupin found in plants (>30 copies in Arabidopsis). Despite the extensive amount of biochemical information about the archetypal cupin, germin from wheat and barley, its enzyme function, namely as an oxalate oxidase (EC 1.2.3.4) that catalyzes the manganese-dependent oxidative decarboxylation of oxalate to carbon dioxide and hydrogen peroxide, was discovered relatively recently. Subsequently, native germin (Woo et al., 2000) and specific members of the related GLPs from both moss and higher plants were shown to have superoxide dismutase (SOD) activity, a means of deactivating superoxide radicals often associated with growth in stress conditions. For example, results from the moss *Barbula unguiculata* suggest that the induction of an SOD form of GLP by salt stress is caused by dissociation of GLP protein from the cell wall into the medium during the logarithmic phase of growth (Nakata et al., 2002). In a related study, the major nectar protein (Nectarin I) from ornamental tobacco (*Nicotiana langsdorffii*×*Nicotiana sanderae*) was identified as a SOD form of GLP that functions to generate high levels of hydrogen peroxide in nectar, and thereby to serve as a defence against potential pathogens (Thornburg et al., 2003). Indeed there are several such instances of a link between GLP expression and plant defence. For example, one GLP was found to be expressed in microstems of chestnut (*Castanea sativa*) infected with chestnut blight fungus (*Cryphonectria parasitica*) (Schafleitner and Wilhelm, 2002), and two GLPs were amongst 1057 differentially expressed genes in Arabidopsis leaves infected with *Peronospora parasitica* (Mahalingam et al., 2003). Similarly, candidate genes involved in both recognition (resistance gene analogues) and general plant defence (putative defence response) in rice were used as molecular markers to test for association with resistance to blast, bacterial blight, sheath blight, and brown plant-hopper (Ramalingam et al., 2003). Blast QTLs were shown to be associated with a GLP together with aldose reductase, JAMyb (a jasmonic acid-induced Myb transcription factor), and peroxidase markers. The latest information of this type is that provided by Segarra et al. (2003) who found that a GLP in the wheat apoplast possesses a serine protease inhibitory activity and thus might be part of a defence system against insect and bacterial proteases.

A similar practical example of the significance of GLP expression in plant development is that in which an in

vitro germination system that reflects relative field emergence potential was used to screen for germination-enhancing and stress-induced genes from seedlings from two varieties of sugar beet (De Los Reyes and McGroth, 2003). Three full-length GLP gene classes were recovered from stress-germinated seedlings of a superior emerging variety. Expression of these genes, together with oxalate oxidase enzyme activity, and the H_2O_2 content of stressed seedlings, were induced by stress germination conditions in this good emerging hybrid and were not induced in a variety that emerged poorly. It was postulated that a block in oxalate metabolism contributeed to lower germination under stress in the lower emerging variety. Selection for such stress-induced GLP expression, or for down-stream targets, presents the first direct target to enable breeding for improved field emergence of sugar beet. An associated example is that of introducing germin as a means of protecting plants against the toxic effects of oxalate-secreting pathogens such as Sclerotinia. This strategy was developed first in transgenic oilseed rape (Thompson et al., 1995) and has been most recently demonstrated in the production of soybean resistant to white mould (Cober et al., 2003).

There are several other recent studies that show interesting patterns of GLP expression without providing evidence of the biochemical function of the proteins involved. For example, in a study of somatic embryogenesis in *Pinus radiata* (Bishop-Hurley et al., 2003) a small number of genes were shown to have high mRNA transcript levels in embryogenic tissue and little or no expression in non-embryogenic (roots, shoots and needles) or callus tissue (needle and fibre culture callus). These gene families included four putative extracellular proteins, namely germin, β-expansin, cellulase and a 21 kDa protein precursor. Another related recent study (Kishi et al., unpublished) is that on the isolation of a tobacco GLP (gi|31711507) as a candidate protein localizing in branched plasmodesmata.

2.3. Phosphomannose isomerase (EC 5.3.1.8)

Phosphomannose isomerase (PMI) catalyzes the interconversion of D-fructose-6-phosphate (F6P) and D-mannose-6-phosphate (M6P). It plays a critical role in the metabolism of D-mannose and supply of GDP-D-mannose, which is necessary for mannosylation of various structures such as lipopolysaccharides (LPS) and glycoproteins. In addition, GDP-D-mannose is a precursor for other activated sugar nucleotides including GDP-L-fucose, GDP-colitose, GDP-perosamine and GDP-D-rhamnose, which are also involved in the synthesis of various glycoconjugates. Of the three types of PMI that have been defined, type I includes all the eukaryotic PMIs identified so far and the enzymes from *Salmonella typhimurium* and *Escherichia coli*, whereas the type II enzymes are bifunctional PMI/GDP-D-mannose pyrophosphorylases found in bacteria such as *Pseudomonas aeruginosa*, *Xanthomonas campestris*, and *S. typhimurium*. The crystal structure of *C. albicans* type 1 PMI was the second cupin to be been resolved (after phaseolin) and is shown in Fig. 2b. It contains a large insertion between the two conserved domains (Fig. 1) and has one atom of zinc in the active site, with a Glu residue replacing one of the His ligands in Motif 1 (Fig. 2a). The *E. coli* version of this gene is of particular applied interest as it has been used extensively in plant transformation for the development of an efficient method for selecting transgenic plants without using antibiotics or herbicides (Boscariol et al., 2003). However, the plant PMIs, such as gi|15219821 from Arabidopsis, have not been the subject of any functional analysis.

2.4. Auxin binding protein

For many years efforts have been made to understand the structure–function relationships of ABP, a protein known to be involved in a variety of plant growth responses. Finally, the structure of auxin-binding protein 1 (ABP1) from maize was determined at 1.9 Å resolution, revealing its auxin-binding site (Woo et al., 2002), and confirming it as a cupin. The binding pocket of ABP1 is predominantly hydrophobic with a zinc ion (cf PMI) deep inside the pocket coordinated by three His and one Glu residues. Auxin binds within this pocket, with its carboxylate binding the zinc and its aromatic ring binding hydrophobic residues including Trp151. There is a single disulphide between Cys2 and Cys155. No conformational rearrangement of ABP1 was observed when auxin bound to the protein in the crystal, but examination of the structure reveals a possible mechanism of signal transduction. The similarity of its auxin binding site to the active site of other enzymatic cupins suggests the likelihood that ABP itself also has an enzymatic function.

2.5. Nuclear proteins

The two main groups of nuclear cupins are the centromeric protein CENP-C (Trazzi et al., 2002), a protein not yet identified in plants, and pirin, a ubiquitous protein encoded by a small gene family in plants. Experimental evidence on this latter protein comes from study of programmed cell death (PCD), similar to animal apoptosis, induced in tomato suspension cells by the topoisomerase I inhibitor camptothecin. In this study, a new tomato gene (Le-pirin) (gi|6651244) was isolated, whose mRNA levels dramatically increase during camptothecin-induced PCD (Orzaez et al., 2001). The Le-pirin encoded protein shows 56% identity with the human protein PIRIN, a nuclear factor reported to

interact with the human oncogene Bcl-3. Human pirin stabilizes the formation of quaternary complexes between Bcl-3, the anti-apoptotic transcription factor NF-κB and its DNA target sequences in vitro. The isolation of Le-pirin and its implication in plant PCD provides new clues on the role of putative NF-kB-associated pathways in plant defence mechanisms.

More recently, a yeast two-hybrid screen was performed to search for proteins that interact with Arabidopsis G protein alpha-subunit (GPA1) (Lapik and Kaufman, 2003). One of the identified GPA1-interacting proteins is AtPirin1 (gi|15231656). The GPA1–AtPirin1 interaction was confirmed in an in vitro binding assay. The authors also characterized two *atpirin1* T-DNA insertional mutants and established that they display a set of phenotypes similar to those of *gpa1* mutants, including reduced germination levels in the absence of stratification and an abscisic acid-imposed delay in germination and early seedling development. These data indicate that AtPirin1 likely functions immediately downstream of GPA1 in regulating seed germination and early seedling development.

2.6. Multidomain monocupins

One intriguing example of this type is the so-called damage repair toleration (DRT) protein (eg gi|18397039 from Arabidopsis) (Pang et al., 1993), a two domain protein in which a cupin C-terminal domain is fused to an N-terminal ribose phosphate domain.

3. Bicupins

These proteins are most likely to have evolved from the duplication and then fusion of a single domain ancestor (Dunwell et al., 2000, 2001), though the possibility remains in some instances of the fusion of two different monocupin precursors.

3.1. Dioxygenases

These bicupins usually only contain single active site in one of the two domains, with the other domain remaining as a non-functional vestigial remnant (Fig. 1).

3.1.1. Quercetin dioxygenase (EC 1.13.11.24)

Quercetin 2,3-dioxygenase (2,3QD) is a copper-containing dioxygenase that catalyses the oxidation of the plant flavonol quercetin (5,7,3′,4′-tetrahydroxy flavonol) to 2-protocatechuoylphloroglucinol carboxylic acid with concomitant production of carbon monoxide. Recently, the crystal structure of 2,3QD from *Aspergillus japonicus* was determined at 1.6 Å resolution (Fusetti et al., 2002); the enzyme was shown to form homodimers, which are stabilized by an *N*-linked heptasaccharide at the dimer interface (Fig. 2b). It represents the first example of a copper-containing cupin and the first resolved structure of a copper dioxygenase. The mononuclear type 2 copper centre within the N-terminal domain displays two distinct geometries: a distorted tetrahedral coordination, formed by His66, His68, His112, and a water molecule, and a distorted trigonal bipyramidal environment, which additionally comprises Glu73 (Fig. 1).

In a mechanistic study (Kooter et al., 2002) characterized 2,3QD by electron paramagnetic resonance spectroscopy. Of a variety of flavonoids studied, only flavonols were able to bind to the copper centre of the enzyme; nine flavonols with different hydroxylation patterns at the A- and B-ring were analysed. X-ray absorption spectroscopy has also been used (Steiner et al., 2002) to characterize the local structural environment of the active site. The applied value of this enzyme is as a possible detergent additive to degrade plant-based stains during washing.

3.1.2. Gentisate 1,2-dioxygenase (EC 1.13.11.4), homogentisate dioxygenase (EC 1.13.11.5) and 1-hydroxy-2-naphthoate dioxygenase (EC 1.13.11.38)

These three dioxygenases are iron-containing enzymes involved in the degradation of aromatic compounds. Of the three types, only homogentisate dioxygenase (HGO), an enzyme that cleaves the aromatic ring during the metabolic degradation of Phe and Tyr, is known to occur in eukaryotes, though its role in plants is uncertain. It is also an unusual type of cupin in having the active site metal ion coordinated near the interface between subunits in the HGO trimer by a Glu and two His side chains.

3.2. Oxalate decarboxylase (EC 4.1.1.2)

Oxalate decarboxylase (OXDC) is a manganese-dependent bicupin enzyme that catalyzes the conversion of oxalate to formate and carbon dioxide. Originally thought to be restricted to fungi, particularly wood-rotting fungi in which oxalate helps activate lignin-degrading enzymes, OXDC has now been identified in several bacteria and recently the structure of one of the two paralogues from *Bacillus subtilis* was determined (Anand et al., 2002). Each of the two cupin domains within the hexameric enzyme (Fig. 2b) contains one manganese-binding site, with the same four manganese-binding residues (three His and one Glu) as present in germin (Fig. 1) (see above). In a mechanistic study using heavy-atom kinetic isotope effect measurements, a multistep model was proposed in which a reversible, proton-coupled, electron transfer from bound oxalate to the Mn-enzyme gives an oxalate radical, which decarboxylates to yield a formate radical anion (Reinhardt et al., 2003). Subsequent reduction and protonation of this intermediate then gives formate.

3.3. Seed storage globulins

Of all the many and varied proteins found within seeds, the most widely distributed are the globular proteins known as legumins and vicilins (Dunwell, 2003). These occur not only in dicots and monocots (including cereals and palms) but also in fern spores. It was the resolution of the vicilin phaseolin from *Phaseolus vulgaris* (Lawrence et al., 1994) that lead indirectly to the discovery of the cupin superfamily, when its structure was used correctly to predict the active site of germin (Gane et al., 1998); the final resolution of the structure of germin (Woo et al., 2000) confirmed that the globulins were indeed simply deactivated enzymes from which the metal binding ligands had been lost during evolution. For example, inspection of the canavalin sequence (Fig. 1) compared with that of germin shows the loss of three of the four metal-binding His and Glu residues in both of the two domains. Presumably it was the thermotolerance of the cupin tertiary structure that lead to its being coopted by plants as a means of storing amino acids, first in the desiccated spore and then in seeds after the evolution of angiosperms.

The additional importance of knowledge about the tertiary and quaternary structures of these globulins lies in their nutritional importance and particularly in their allergenic potential since many of the individual allergenic epitopes have now been mapped onto the structure (Mills et al., 2002; Xiang et al., 2002). This information provides a potential route to the removal of such epitopes using recombinant techniques (Dunwell, 2002; Maruyama et al., 2001).

3.4. Storage protein/sucrose binding protein

A form of bicupin related to the seed globulins is the sucrose binding protein (SBP), a protein implicated as an important component of the sucrose uptake system in plants, particularly legumes. SBP-mediated sucrose transport displays unique kinetic features and the protein is not similar to other transport proteins. Amongst recent studies is one (Pirovani et al., 2002) reporting the characterization of a membrane-associated SBP protein from soybean (*Glycine max*) that exhibits GTP binding activity. A site directed mutant failed to bind GTP but retained the ability to undergo oligomerization and to promote growth of the susy7 yeast strain, deficient in utilizing extracellular sucrose, on medium containing sucrose as the sole carbon source. These results indicate that GTP binding and sucrose transport by this SBP are separable and function independently.

4. Multicupins

Recent analysis of genomic data has identified several examples of proteins that contain multiple (>2) copies of cupin domains (Dunwell, unpublished). These include potential examples such as the *Arabidopsis* protein gi|25406927 that comprises four 2OG-Fe^{2+} domains; to date it has no assigned function.

5. Concluding remarks

This review has described the remarkable chemical diversity found in a single protein superfamily, and confirmed the cupins as being amongst the most versatile of all protein folds described to date (Anantharanan et al., 2003). This realisation has come largely from the recent ability efficiently to integrate information from genomic sequencing projects with the great increase in the number of 3D protein structures (Table 1). Without this integration it would not have been possible to undercover the distant similarities between the various cupin members (Khuri et al., 2001), and to trace an evolutionary path that has lead from small microbial enzymes (e.g. the 89 AA protein gi|33357290 from the thermophilic bacterium *Thermotoga maritima*, 1LKN) to the multidomain transcription factors such as members of the jmjC family (Clissold and Ponting, 2001), and the trimeric/hexameric seed storage globulins (400–500 AA in each monomer) (Dunwell, 2003) that form a vital part of the human diet.

Acknowledgements

Part of the work described here was supported by grants from the Biotechnology and Biological Sciences Research Council (BBSRC).

References

Aas, P.A., Otterlei, M., Falnes, P.O., Vagbo, C.B., Skorpen, F., Akbari, M., Sundheim, O., Bjoras, M., Slupphaug, G., Seeberg, E., Krokan, H.E., 2003. Human and bacterial oxidative demethylases repair alkylation damage in both RNA and DNA. Nature 421, 859–863.

Anand, R., Dorrestein, P.C., Kinsland, C., Begley, T.P., Ealick, S.E., 2002. Structure of oxalate decarboxylase from *Bacillus subtilis* at 1.75 Å resolution. Biochemistry 41, 7659–7669.

Anantharanan, V., Aravind, L., Koonin, E.V., 2003. Emergence of diverse biochemical activities in evolutionarily conserved scaffolds of proteins. Curr. Opin. Chem. Biol. 7, 12–20.

Aravind, L., Koonin, E.V., 2001. The DNA-repair protein AlkB, EGL-9, and leprecan define new families of 2-oxoglutarate- and iron-dependent dioxygenases. Genome Biol. 2, Research 7.1-7, 8.

Begly, T.J., Samson, L.D., 2003. AlkB mystery solved: oxidative demethylation of N1-methyladenine and N3-methylcytosine adducts by a direct reversal process. Trends Biochem. Sci. 28, 2–5.

Bishop-Hurley, S.L., Gardner, R.C., Walter, C., 2003. Isolation and molecular characterization of genes expressed during somatic embryo development in *Pinus radiata*. Plant Cell Tissue Organ Cult 74, 267–281.

Boscariol, R.L., Almeida, W.A., Derbyshire, M.T., Mourao Filho, F.A., Mendes, B.M. 2003. The use of the PMI/mannose selection system to recover transgenic sweet orange plants (*Citrus sinensis* L. Osbeck). Plant Cell Rep. (Epub ahead of print).

Clissold, P.M., Ponting, C.P., 2001. JmjC: cupin metalloenzyme-like domains in jumonji, hairless and phospholipase $A_2\beta$. Trends Biochem. Sci. 26, 7–9.

Cober, E.R., Rioux, S., Rajcan, I., Donaldson, P.A., Simmonds, D.H., 2003. Partial resistance to white mold in a transgenic soybean line. Crop Sci. 43, 92–95.

De Los Reyes, B.G., McGrath, J.M., 2003. Cultivar-specific seedling vigor and expression of a putative oxalate oxidase germin-like protein in sugar beet (*Beta vulgaris* L.). Theor. Appl. Genet. 107, 54–61.

Dunwell, J.M., 1998. Cupins: a new superfamily of functionally-diverse proteins that include germins and plant seed storage proteins. Biotechnol. Genet. Engin. Rev. 15, 1–32.

Dunwell, J.M., 2002. Future prospects for transgenic crops. Phytochem. Rev. 1, 1–12.

Dunwell, J.M., 2003. Structure, function and evolution of the legumin seed storage proteins. In: Steinbüchel, A., Fahnestock, S.R. (Eds.), Biopolymers, Vol. 8. Polyamides and Complex Proteinaceous Materials II. Wiley-VCH, Weinheim, pp. 223–253.

Dunwell, J.M., Culham, A., Carter, C.E., Sosa-Aguirre, C.R., Goodenough, P.W., 2001. Evolution of functional diversity in the cupin superfamily. Trends Biochem. Sci. 26, 740–745.

Dunwell, J.M., Gane, P.J., 1998. Microbial relatives of seed storage proteins: conservation of motifs in a functionally diverse superfamily of enzymes. J. Mol. Evol. 46, 147–154.

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

Falnes, P.O., Johansen, R.F., Seeberg, E., 2002. AlkB-mediated oxidative demethylation reverses DNA damage in *Escherichia coli*. Nature 419, 178–182.

Frey, M., Huber, K., Park, W.J., Sicker, D., Lindberg, P., Meeley, R.B., Simmons, C.R., Yalpani, N., Gierl, A., 2003. A 2-oxoglutarate-dependent dioxygenase is integrated in DIMBOA-biosynthesis. Phytochem. 62, 371–376.

Gane, P.J., Dunwell, J.M., Warwicker, J., 1998. Homology modelling based on the structure of vicilins predicts a histidine cluster in the active site of oxalate oxidase. J. Mol. Evol. 46, 488–493.

Hewitson, K.S., McNeill, L.A., Riordan, M.V., Tian, Y.M., Bullock, A.N., Welford, R.W., Elkins, J.M., Oldham, N.J., Bhattacharya, S., Gleadle, J.M., Ratcliffe, P.J., Pugh, C.W., Schofield, C.J., 2002. Hypoxia-inducible factor (HIF) asparagine hydroxylase is identical to factor inhibiting HIF (FIH) and is related to the cupin structural family. J. Biol. Chem. 277, 26351–26355.

Fusetti, F., Schroter, K.H., Steiner, R.A., van Noort, P.I., Pijning, T., Rozeboom, H.J., Kalk, K.H., Egmond, M.R., Dijkstra, B.W., 2002. Crystal structure of the copper-containing quercetin 2,3-dioxygenase from *Aspergillus japonicus*. Structure 10, 259–268.

Khuri, S., Bakker, F.T., Dunwell, J.M., 2001. Phylogeny, function and evolution of the cupins, a structurally conserved, functionally diverse superfamily of proteins. Mol. Biol. Evol. 18, 593–605.

Kooter, I.M., Steiner, R.A., Dijkstra, B.W., van Noort, P.I., Egmond, M.R., Huber, M., 2002. EPR characterization of the mononuclear Cu-containing *Aspergillus japonicus* quercetin 2,3-dioxygenase reveals dramatic changes upon anaerobic binding of substrates. Eur. J. Biochem. 269, 2971–2979.

Lapik, Y.R., Kaufman, L.S., 2003. The Arabidopsis cupin domain protein atpirin1 interacts with the G protein alpha-subunit GPA1 and regulates seed germination and early seedling development. Plant Cell 15, 1578–1590.

Lawrence, M.C., Izard, T., Beuchat, M., Blagrove, R.J., Colman, P.M., 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.

Mahalingam, R., Gomez-Buitrago, A.M., Eckardt, N., Shah, N., Guevara-Garcia, A., Day, P., Raina, R., Fedoroff, N.V., 2003. Characterizing the stress/defense transcriptome of *Arabidopsis*. Genome Biol. 4, R20.

Maruyama, N., Adachi, M., Takahashi, K., Yagasaki, K., Kohno, M., Takenaka, Y., Okuda, E., Nakagawa, S., Mikami, B., Utsumi, S., 2001. Crystal structures of recombinant and native soybean beta-conglycinin beta homotrimers. Eur. J. Biochem. 268, 3595–3604.

Mills, E.N.C., Jenkins, J., Marigheto, N., Belton, P.S., Gunning, A.P., Morris, V.J., 2002. Allergens of the cupin superfamily. Biochem. Soc. Trans. 30, 925–929.

Monna, L., Kitazawa, N., Yoshino, R., Suzuki, J., Masuda, H., Maehara, Y., Tanji, M., Sato, M., Nasu, S., Minobe, Y., 2002. Positional cloning of rice semidwarfing gene, sd-1: rice "green revolution gene" encodes a mutant enzyme involved in gibberellin synthesis. DNA Res. 9, 11–17.

Nakata, M., Shiono, T., Watanabe, Y., Satoh, T., 2002. Salt stress-induced dissociation from cells of a germin-like protein with Mn-SOD activity and an increase in its mRNA in a moss, *Barbula unguiculata*. Plant Cell Physiol. 43, 1568–1574.

Orzaez, D., de Jong, A.J., Woltering, E.J., 2001. A tomato homologue of the human protein PIRIN is induced during programmed cell death. Plant Mol. Biol. 46, 459–468.

Pang, Q., Hays, J.B., Rajagopal, I., Schaefer, T.S., 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. J. Plant Mol. Biol. 22, 411–426.

Pirovani, C.P., Macedo, J.N., Contim, L.A., Matrangolo, F.S., Loureiro, M.E., Fontes, E.P., 2002. A sucrose-binding protein homologue from soybean exhibits GTP-binding activity that functions independently of sucrose transport activity. Eur. J. Biochem. 269, 3998–4008.

Pochapsky, T.C., Pochapsky, S.S., Ju, T., Mo, H., Al-Mjeni, F., Maroney, M.J., 2002. Modeling and experiment yields the structure of acireductone dioxygenase from *Klebsiella pneumoniae*. Nat. Struct. Biol. 9, 966–972.

Ramalingam, J., Vera Cruz, C.M., Kukreja, K., Chittoor, J.M., Wu, J.L., Lee, S.W., Baraoidan, M., George, M.L., Cohen, M.B., Hulbert, S.H., Leach, J.E., Leung, H., 2003. Candidate defense genes from rice, barley, and maize and their association with quantitative resistance in rice. Mol. Plant Microbe Interact. 16, 14–24.

Reinhardt, L.A., Svedruzic, D., Chang, C.H., Cleland, W.W., Richards, N.G.J., 2003. Heavy isotope effects on the reaction catalyzed by the oxalate decarboxylase from *Bacillus subtilis*. J. Am. Chem. Soc. 125, 1244–1252.

Segarra, C.I., Casalongué, C.A., Pinedo, M.L., Ronchi, V.P., Conde, R.D., 2003. A germin-like protein of wheat leaf apoplast inhibits serine proteases. J. Exp. Bot. 54, 1335–1341.

Spielmeyer, W., Ellis, M.H., Chandler, P.M., 2002. Semidwarf (sd-1), "green revolution" rice, contains a defective gibberellin 20-oxidase gene. Proc. Natl. Acad. Sci. U.S.A. 99, 9043–9048.

Schafleitner, R., Wilhelm, E., 2002. Isolation of wound-responsive genes from chestnut (*Castanea sativa*) microstems by mRNA display and their differential expression upon wounding and infection with the chestnut blight fungus (*Cryphonectria parasitica*). J. Physiol. Mol. Plant Pathol. 61, 339–348.

Steiner, R.A., Kalk, K.H., Dijkstra, B.W., 2002. Anaerobic enzyme substrate structures provide insight into the reaction mechanism of the copper-dependent quercetin 2,3-dioxygenase. Proc. Natl. Acad. Sci. U.S.A. 99, 16625–16630.

Thompson, C., Dunwell, J.M., Johnstone, C.E., Lay, V., Ray, J., Schmitt, M., Watson, H., Nisbet, G., 1995. Degradation of oxalic-acid by transgenic oilseed rape plants expressing oxalate oxidase. Euphytica 85, 169–172.

Thornburg, R.W., Carter, C., Powell, A., Mittler, R., Rizhsky, L., Horner, H.T., 2003. A major function of the tobacco floral nectary is defense against microbial attack. Plant System. Evolut. 238, 211–218.

Trazzi, S., Bernardoni, R., Diolaiti, D., Politi, V., Earnshaw, W.C., Perini, G., Valle, G.D., 2002. In vivo functional dissection of human inner kinetochore protein CENP-C. J. Struct. Biol. 140, 39–48.

Trewick, S.C., Henshaw, T.F., Hausinger, R.P., Lindahl, T., Sedgwick, B., 2002. Oxidative demethylation by *Escherichia coli* AlkB directly reverts DNA base damage. Nature 419, 174–178.

Wilmouth, R.C., Turnbull, J.J., Welford, R.W., Clifton, I.J., Prescott, A.G., Schofield, C.J., 2002. Structure and mechanism of anthocyanidin synthase from *Arabidopsis thaliana*. Structure 10, 93–103.

Woo, E.J., Marshall, J., Bauly, J., Chen, J.G., Venis, M., Napier, R.M., Pickersgill, R.W., 2002. Crystal structure of auxin-binding protein 1 in complex with auxin. EMBO J. 21, 2877–2885.

Woo, E.J., Dunwell, J.M., Goodenough, P.W., Marvier, A.C., Pickersgill, R.W., 2000. Germin is a manganese containing homohexamer with oxalate oxidase and superoxide dismutase activities. Nat. Struct. Biol. 7, 1036–1040.

Xiang, P., Beardslee, T.A., Zeece, M.G., Markwell, J., Sarath, G., 2002. Identification and analysis of a conserved immunoglobulin E-binding epitope in soybean G1a and G2a and peanut Ara h 3 glycinins. Archiv. Biochem. Biophys. 408, 51–57.