

***Arabidopsis thaliana* contains a large family of germin-like proteins: characterization of cDNA and genomic sequences encoding 12 unique family members**

Clay Carter, Richard A. Graham and Robert W. Thornburg*

2212 Molecular Biology Building, Department of Biochemistry and Biophysics, Iowa State University Ames, IA 50011, USA (*author for correspondence)

Received 17 March 1998; accepted in revised form 25 April 1998

Key words: germin, germin-like proteins, oxalate oxidase, multigene family, *Arabidopsis thaliana*

Abstract

We have identified 39 *Arabidopsis thaliana* ESTs encoding germin-like proteins (GLPs) and have completely sequenced 25 of these cDNAs. Our analysis demonstrates that the *Arabidopsis* genome contains a gene family with at least 12 GLP genes. Comparisons with other known germins and germin-like proteins indicate that these *Arabidopsis* GLP subfamilies are unique from wheat germin. All other known GLPs fall into one of these subfamilies. The translated GLPs show approximately 35% amino acid identity with other GLPs outside of their subfamily and significantly higher levels of identity within their respective subfamily. The 3' ends of many of the GLP cDNAs are heterogeneous and several sites of polyadenylation are used. Ten of the GLPs have N-terminal signal sequences and most appear to be exported from the cell. Structurally, the GLPs are predicted to have a high content of β -pleated sheet. Seven conserved regions of β -sheet were found in each of the GLP proteins along with α -helices located at both N- and C-termini. These same structural elements are also conserved in wheat germin. With one exception, all GLP family members contain at least one N-glycosylation site. All of these sites are conserved in an unstructured loop between β -1 and β -2. Genes for two of these GLPs were identified in genomic sequences previously deposited in the GenBank. The GLP3b gene is physically linked to the polyubiquitin 4 gene. The 3' end of the GLP3b mRNA is only 0.5 kb from the *ubq4* start of transcription. Analysis of the GLP3b promoter shows the presence of a single putative auxin-response sequence located at –124 to –111 upstream from the 5' end of the GLP3b mRNA. The GLP9 gene was identified in an *Arabidopsis* contig from Chromosome 4.

Abbreviations: GLPs, germin-like proteins; ABRC, *Arabidopsis* Biological Resource Center at The Ohio State University (arabidopsis + @osu.edu).

Introduction

Germin¹ is a 130 kDa homopentameric protein [43] first detected in germinating cereals. Later, this protein was found to be present in cereal cell walls, and still later by combined analysis of its genetic coding elements and assay of its possible activity was found to be an enzyme: oxalate oxidase (for review, see [38]).

Germin is expressed primarily in germinating embryos of monocots.

Proteins with sequence identity to germins have been identified from wheat [37] as well as from other plant species. The expression of these germin-like proteins (GLPs) varies widely among plant species. In *Sinapis alba*, a germin-like protein, SaGLP, is expressed in a circadian oscillation in the epidermis and spongy parenchyma of young leaves [26]. This accumulation occurs in the extracellular spaces and when the primary cell wall material is lost, the SaGLP is

¹ Journal Paper No. J-17352 of the Iowa Agriculture and Home Economics Experiment Station, Ames, Iowa. Project No. 3340.

no longer detected. In the ice plant, *Mesembryantum crystallinum*, a germin-like protein (McGLP) is expressed in roots and decreases in response to salt stress [47]. The sequence of a GLP from Japanese Morning Glory, *Pharbitis nil* (PnGLP), has also been deposited in the GenBank [50], but no information is available on its expression. Extracellular GLPs that share both immunological identity and sequence identity with wheat germin have also been identified in embryogenic cell cultures of Caribbean pine, *Pinus caribaea* [16]. These proteins are not expressed in nonembryogenic cell cultures.

Recently, germin has been identified as an oxalate oxidase [19, 38]. Oxalate oxidase catalyzes the oxidative breakdown of oxalate ($C_2H_2O_4$) into 2 CO_2 plus H_2O_2 . It has been suggested that germin contributes to cell wall restructuring [38, 60] by producing H_2O_2 which is required for the peroxidase catalyzed cross-linking of many cell wall components [21, 56].

An alternative recent proposal for the function of germin and germin-like proteins is that these proteins may be involved in plant defense [34, 33]. Hydrogen peroxide has been shown to play a central role in plant defenses. H_2O_2 has direct antimicrobial effects [34], and has also been proposed to function as a second messenger in the activation of defense gene expression, including phytoalexin biosynthetic pathway, hypersensitive responses and the PR responses [3, 13, 14, 39]. Further, H_2O_2 participates in oxidative crosslinking of structural cell wall proteins [9, 10, 56] and induces specific genes encoding enzymes such as glutathione S-transferase that are involved in plant protection [39].

Some experimental evidence supporting the role of germin-like proteins in plant defense already exists. Indeed, a possible role for germins in plant defense was suggested prior to the identification of their enzymatic activity. Germins are conspicuous in their presence in rust-resistant, domesticated wheat and other cereal crops [35]. Recently, it was independently shown by several groups that both germin and oxalate oxidase activity increase during powdery mildew attack of barley [20, 28, 60]. Whether this increase in the germin-like proteins of barley in response to powdery mildew is a result of a specific pathogen mediated induction or a response to altered osmotic state is not yet clear. Nevertheless, because the *Arabidopsis* GLPs have not been associated with a known enzyme activity, the association between these proteins and plant defense remains tenuous.

Recently, it has become clear that multiple forms of germin-like proteins exist in plants. Three germin-like proteins were recently identified from *Arabidopsis* [46]. These proteins were termed Germin 1, Germin 2, and Germin 3; however, our analysis of the *Arabidopsis* germin-like protein gene family indicates that these proteins are not true germins and we suggest that they be termed germin-like proteins rather than germins. We have undertaken this characterization of the *Arabidopsis* GLP gene families to establish a concise framework of all *Arabidopsis* GLPs for the further study.

Materials and methods

cDNA sequencing

DNA sequence reactions were performed using the Applied Biosystems Prism Dye-deoxy Cycle Sequencing Kit. The reactions were run on an Applied Biosystems Prism 377 DNA sequencer, Perkin-Elmer Corp. Sequence was initiated from known vector sequences. On the basis of these runs, primers specific to each GLP sequence were constructed to extend the DNA sequence. DNA sequences were performed in duplicate or triplicate for each run. Each strand was completely sequenced, including sequencing through all restriction sites and the entire sequence of each cDNA was confirmed on the opposite strand.

DNA analysis tools

EST database searches were conducted using several on-line tools, including: the University of Minnesota search engines of EST Analysis Files using WAIS (http://lenti.med.umn.edu/general_cdna/wais_search.html) and BLASTn ([2] at <http://lenti.med.umn.edu/cgi-bin/blast/blastn.cgi>), the TIGR database *Arabidopsis* EST name searcher (http://www.tigr.org/tdb/at/-searching_at/at_name_search/at_name_search.html), the Agricultural Genome Information Server database search engine (<http://probe.nalusda.gov:8300/cgi-bin/query>) and the NCBI dbEST search engine (<http://www.ncbi.nlm.nih.gov/dbEST/index.html>).

The identity of each completed cDNA sequence was verified using BLAST homology searches of the GenBank [2]. The Wisconsin GCG package [22] and DNA Strider [41] were used for a variety of analyses on the DNA and protein sequences. The N-terminal signal sequences were identified using the PSORT protein sorting prediction tool ([49] at

<http://psort.nibb.ac.jp/>). The molecular mass and the pI were computed with the ExPASy Compute pI/Mw tool ([8] at http://expasy.hcuge.ch/ch2d/pi_tool.html). The secondary structure was analyzed with the Protein Predict tool at EMBL in Heidelberg ([52, 53] at <http://www.embl-heidelberg.de/predictprotein/predictprotein.html>).

Results

Arabidopsis contains a large multigene family of GLPs

To identify already cloned germin-like proteins, we searched the *Arabidopsis* EST GenBank submissions for germin-like proteins. These ESTs have been isolated from 16 different cDNA libraries prepared for the *Arabidopsis* genome project in the USA and in France [1].

This search identified 39 cDNAs encoding GLPs from 35,680 total cDNAs in the *Arabidopsis* EST databases. Thus, mRNAs encoding GLPs represent about 0.1% of all of the mRNAs cloned in these libraries. All of these ESTs were obtained from ABRC and 25 of these clones were completely sequenced. Towards the end of our analyses, nine clones (primarily GLP1s, GLP3bs and GLP5s) were assigned to their respective gene families after partial sequencing. The identity and GenBank accession numbers of each of the cDNAs used in this study are shown in Table 1.

In addition, *Arabidopsis* cDNAs encoding three germin-like proteins termed At-Germ1 [51], At-Germ2 [25], and At-Germ3 [46] have been deposited in GenBank. These cDNAs have been included in our analysis of the germin-like proteins from *Arabidopsis*. One GLP1 cDNA apparently has been lost from the ABRC collection [23].

Expression of the GLP cDNAs

Alignment of the DNA sequences of the 39 sequenced GLP cDNAs identified at least 12 unique sequences that are expressed from the *Arabidopsis thaliana* genome. The sequences GLP1, GLP2, and GLP3 correspond to the earlier published Germin 1, Germin 2, and Germin 3 sequences [25, 46, 51]. The remaining sequences were named GLP4 through GLP10 in the order in which they were sequenced. Two of these, GLP2 and GLP3, have duplicate sequences that are nearly identical and are named GLP2a/2b and GLP3a/3b.

The frequencies that each GLP clone was identified are also presented in Table 1. Of all *Arabidopsis* GLP cDNAs, 36% (14/39) encoded the GLP1 protein. The second most abundant GLP cDNA identified was GLP3b (6/39) followed by GLP5 (5/39), GLP2a (3/39), GLP3a (2/39) and GLP10 (2/39). The remainder of the GLPs (GLP2b, GLP4, GLP6, GLP7, GLP8 and GLP9) were identified from single clones. Thus, based upon the frequency of cloned transcripts it appears that the GLP1 gene is expressed at 2.5 to 14 times higher levels than other GLP genes.

The majority of these ESTs were isolated from libraries made of mRNA from pooled tissues; consequently, little information is available on their expression. Some clones, however, were isolated from tissue specific libraries, and for these clones we can identify the tissues where the genes are expressed. As is shown in Table 2, for the most highly expressed GLP, GLP1, several clones were isolated from young shoots and whole seedlings. GLP1 clones were also found in libraries constructed from etiolated as well as green tissues. For the second most abundant GLPs, GLP3b and GLP5, several clones were isolated from the same library derived from 3-day old seedling hypocotyls. One GLP3b clone was isolated from a cDNA library prepared from sliced leaves that were incubated in liquid culture. It is thus interesting to note that for those clones where a tissue of identity can be assigned, the vast majority come from young shoots, etiolated and green seedlings or from hypocotyl shortly after germination. The only clones that were not present in young tissues were the GLP2a/b clones that were expressed in immature siliques. In every case, these GLP2 clones, representing 10% of all the GLP clones isolated, were isolated from siliques. The remainder of the clones, GLPs 3a, 4, 6, 7, 8, 9, and 10, representing 26% of the isolates, could not be assigned to any tissues of origin.

The GLP cDNA family

To evaluate the relationships between these cDNAs, the nucleotide sequences were aligned with the GCG tool, 'PileUp' [22]. This program creates a multiple sequence analysis using progressive pairwise alignments. It also generated the dendrogram shown in Figure 1A. This figure shows the relationships between all published germin and GLP cDNAs. Of all GLP cDNA sequences analyzed, GLP7 is the most unique. This clone does not appear to have any close relatives among any of the *Arabidopsis* GLPs examined or

Table 1. Classification of germin-like proteins from *Arabidopsis*.

Number of clones found	GLP locus	GenBank accession numbers		EST clone name
		this work ^a	previous	
14	GLP1	U75189	H36759	178G21T7
		U75190	T13617	21C9T7
		U75196	Z30804/5	43546/7 VBVC10 ^e
		U75197	N38578	220E6T7
		U75201	T22370	104E20T7
		U75206	T22353	104D20T7
		U95034	W43261	249C3T7
		U95035	T14124	47F11T7
		b	T88402	156B13T7
		b	Z18183	14032
		b,c	N38403	216C14T7
		b,d	X91921	(At-GERM1)
		b	AA067403	88L19T7
		b	AA067568	85C7T7
3	GLP2a	U75192	Z17674	14147 (YAP134T)
		U75204	Z17644	14105 (YAP091T7)
		b	AA395913	304B1T7
1	GLP2b	b,d	X91957	(At-GERM2)
2	GLP3a	U75188	T44248	111F21T7
		U75203	T14021	45D6T7
6	GLP3b	U75193	T44629	128A6T7
		U75195	T42098	110H21T7
		U75205	H76449	195B22T7
		b	Z26437	(At-GERM3)
		b	AA042754	E8E6T7
		b	AA040958	E2F12T7
1	GLP4	U75187	T42670	114H1T7
6	GLP5	U75191	H36918	180L10T7
		U75198	N38123	219H4T7
		U75199	N38257	222A11T7
		U75200	R65243	168B5T7
		b	T46829	145C14T7
		b	AA041156	E2C8T7
1	GLP6	U75194	R84146	157N14T7
1	GLP7	U75202	T88481	157B12T7
1	GLP8	U75207	T44499	125O15T7
1	GLP9	U81294	N65240	223N10T7
2	GLP10	U95036	W43342	250C6T7
		b	AA042591	269F10T7

^aSequences of clones used in this work were deposited in GenBank as new accessions with links to the older accessions that were only partially sequenced.

^bThese clones were partially sequenced. The resulting sequence was sufficient to assign these clones to their respective families.

^cGenBank locus N38403 contains the sequence of a germin-like protein, however, we requested this clone twice from ABRC. Both times the clone supplied was of a different cDNA related to *Zea mays* IAA glucose synthase (GenBank U81293) [23]. Apparently the clone having the sequence reported in GenBank locus N38403 has been lost.

^dThe *Arabidopsis thaliana* clone for At-Germ1, At-Germ2, and At-Germ3 were isolated and characterized by Dr M. Raynal and coworkers (Université de Perpignan, France)

^eClones 4346 (Atts2554) and 43547 (Atts2555) are the same clone sequenced from either end.

Table 2. Tissue of cDNA origin

Protein	% of Clones	Expression
GLP1	36	etiolated seedlings whole seedlings green shoots
GLP 2a/b	10	immature siliques
GLP3b	15	seedling hypocotyl, 3-day sliced leaves in liquid culture
GLP5	13	seedling hypocotyl, 3-day
all others	26	mixed tissues

among any previously characterized germin-like proteins. The remainder of the GLPs belong to one of four subfamilies. The wheat and barley germis form one subfamily (the true germis) that is dissimilar from all other germin-like protein subfamilies. None of the *Arabidopsis* GLPs are closely related to the true wheat germis. The GLP subfamily 1 consists of the *Arabidopsis* GLPs (GLP2a, GLP2b, GLP6 and GLP9) and the *Mesembryanthemum crystallinum* germin-like protein, McGLP. The GLP subfamily 2 consists of the *Arabidopsis* GLPs (GLP4, GLP5, GLP8 and GLP10). No non-*Arabidopsis* members of this GLP subfamily 2 have been identified to date from any plant species other than *Arabidopsis*. Finally the GLP subfamily 3, which is the most abundant (comprising 56% of all *Arabidopsis* GLP cDNAs), consists of the *Arabidopsis* GLPs (GLP1, GLP3a, and GLP3b) as well as the GLPs isolated from *Pharbitis nil*, *Brassica napus* and *Sinapis alba*. The BnGLP cDNA is most like the *Arabidopsis* GLP1 cDNA while the SaGLP cDNA is most similar to the GLP3 cDNAs. The PnGLP is the most dissimilar of the GLP subfamily 3 clones; however, based upon the conservation of methionine at position 105 (see below), this clone is a GLP1 member.

The majority of the GLP cDNAs are near full-length with 5'-untranslated regions between 30 and 50 nucleotides long. Exceptions are GLP7, GLP9 and GLP10, each of which is incomplete. The GLP7 clone is missing ca. 200 to 250 nucleotides from the 5' end resulting in about 70 amino acids missing from the N-terminal sequence of the protein. Nevertheless, the remaining sequence of GLP7 is sufficiently distinct from the other GLPs, indicating that it is a novel GLP. The

GLP9 cDNA is missing its 5'-untranslated region and at least one or possibly more codons at the N-terminus of the GLP9 protein. The GLP10 cDNA is missing its 5'-untranslated region and approximately 15 amino acids at the N-terminus of the GLP10 protein.

With the exception of GLP8, none of the full-length cDNAs had the optimal translation initiation sequences identified by Kozak [32]. For those GLP families in which multiple cDNAs were sequenced, we found that the 3' polyadenylations sites were very heterogeneous with utilized sites spread over as much as 100 nucleotides (data not shown).

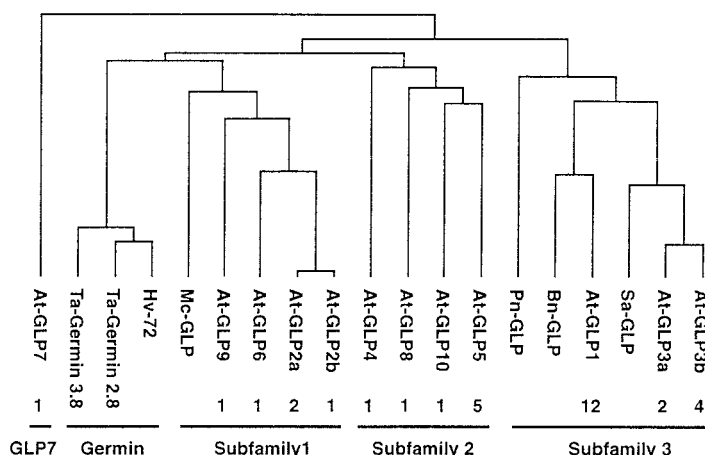
Closely related GLPs

We have identified two GLP members (GLP2 and GLP3) that each have distinct but very closely related multiple cDNAs. We have therefore termed these GLP2a/b and GLP3a/b. We characterized 3 members GLP2 clones, all of which belonged to the GLP2a subfamily. We did not characterize any GLP2b clones, but the AtGerm2 cDNA (GenBank X91957) was sufficiently different from the GLP2a clones to indicate a separate gene. We identified eight GLP3 clones, two GLP3 members being GLP3a, and six members consisting of GLP3b.

Differences between GLP2a and GLP2b. There were seven point mutations observed as differences between the GLP2a and GLP2b cDNAs (Table 3). Two of these were transitions and five were transversions. All occurred within the coding region of the cDNAs. In contrast to the GLP3 genes, there were no deletions observed between GLP2a and GLP2b. Of the seven point mutations, six resulted in changes in the amino acid sequences between GLP2a and GLP2b.

Differences between GLP3a and GLP3b. We identified 23 nucleotide sequence differences between the cDNAs encoding GLP3a and GLP3b. These were scattered throughout the cDNAs (Table 3). Fifteen of these differences were point mutations and the remaining eight were deletions. Fourteen of these point mutations occurred within the coding region of the cDNAs. The remaining one occurred just after the stop codon in the 3' untranslated region. No mutations were identified in the 5'-untranslated region. Of the 15 point mutations, transversions occurred at five locations while transitions occurred at the remaining 10 locations. With one exception all of the point mutations are silent. The single non-silent point mutation is an A → C transversion occurring at position 209 of the cDNA resulting

A. GLP cDNA subfamilies



B. Percent amino acid identity between germins and germin-like proteins

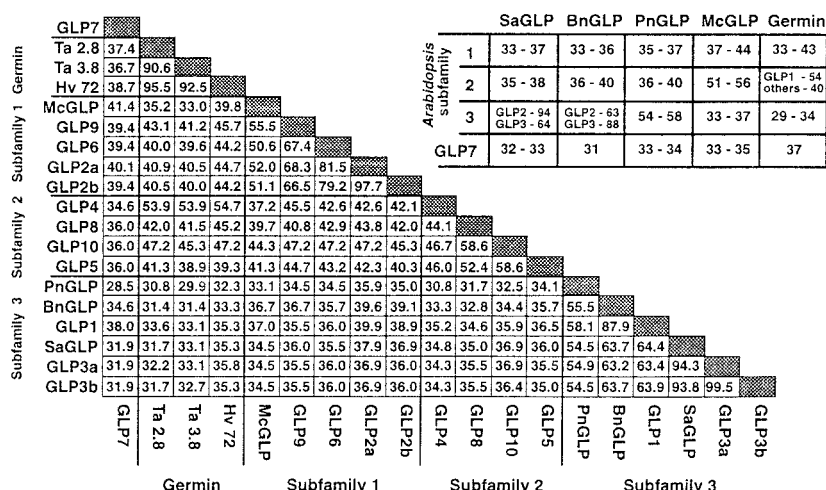


Figure 1. Germin and germin-like protein sequences in plants. A. Relationships of the GLP cDNAs. The dendrogram was prepared with the GCG tool, PileUp. The sequences used for this analysis were At-GLP1 (GenBank Accession U75206); At-GLP2a (U75192); At-GLP2b (X91957); At-GLP3a (U75188); At-GLP3b (U75195); At-GLP4 (U75187); At-GLP5 (U75198); At-GLP6 (U75194); At-GLP7 (U75202); At-GLP8 (U75207); At-GLP9 (U81294); At-GLP10 (U95036); Ta-Germin 2.8, *Triticum aestivum* germin gene 2.8 (M63223, nucleotides 1695–2725); Ta-Germin 3.8, *Triticum aestivum* Germin gene 3.8 (M63224, nucleotides 1217–2238); Hv-72, *Hordeum vulgare* C 72 Germin (U01963); SaGLP, *Sinapis alba* germin-like protein (X84786); McGLP, *Mesembryanthemum crystallinum* germin-like protein (M93041); PnGLP, *Pharbitis nil* germin-like protein (D45425); BnGLP, *Brassica napus* germin-like protein (U21743). The frequency that each of the *Arabidopsis* GLPs were found is presented below their names. The subfamilies of germin-like proteins are GLP7, germin, GLP subfamily 1, GLP subfamily 2, and GLP subfamily 3. B. Amino acid identity between all germins and germin-like proteins. Following alignment of these translated amino acid sequences from the cDNAs and genes shown in A, the sequences were compared to determine the identity between each pair of sequences. The data in the upper right quadrant compares the ranges of identity between the non-*Arabidopsis* germins and GLPs with each of the *Arabidopsis* GLP family groups.

Table 3a. Differences between At-GLP2a and At-GLP2b.

	Position	Type	GLP2a	GLP2B	Mutation
1	196	point	G	C	Gly ⁶³ → Ala ⁶³
2	301	point	G	T	Arg ⁹⁸ → Leu ⁹⁸
3	303	point	A	G	Ile ⁹⁹ → Val ⁹⁹
4	323	point	C	G	Gly ¹⁰⁵ (silent)
5	327	point	A	G	Asn ¹⁰⁷ → Asp ¹⁰⁷
6	412	point	A	T	Gln ¹³⁵ → Leu ¹³⁵
7	413	point	A	T	Gln ¹³⁵ → Leu ¹³⁵

in a Lys-59 to Thr-59 change between GLP3a and GLP3b. All of the deletions occurred within the 3'-untranslated region. Seven of the eight deletions were single nucleotide deletions, and the remaining one was a deletion of 18 nucleotides that is absent in GLP3b but is present in GLP3a.

GLP proteins

The relatedness of the germin-like proteins (deduced from the GLP cDNAs) is shown in Panel B of Figure 1. The data is presented as the percent identity between the various GLPs. All the GLPs share at least 31% identity at the amino acid level. GLP7 shows the lowest level of amino acid identity with the other GLPs (31 to 36%). This is similar to the identity found at the nucleotide level. The wheat and barley germins also share low levels of amino acid identity with the other GLPs. The GLPs in subfamily 1 share 72 to 82% identity with each other (excluding the GLP2a to GLP2b identity, 98%). In general, the *Mesembryanthemum* GLP shares less identity with the other *Arabidopsis* subfamily 1 GLPs than these *Arabidopsis* GLPs share among themselves. The *Arabidopsis* cDNA most like the *Mesembryanthemum* GLP is GLP9. The GLP subfamily 2 shows 46 to 53% identity among themselves, and lower levels of identity with other GLPs. Finally, subfamily 3 shows 62% identity at the amino acid level between GLP1/GLP3a and GLP1/GLP3b. As seen above, the GLP3a and GLP3b share nearly complete (>99.5%) identity. The inset at the upper right of panel B, shows the *Sinapis alba* GLP shares 94% amino acid identity with each of the *Arabidopsis* GLP3 proteins but lower identity (64%) with the GLP1 protein. The *Brassica napus* GLP shares 88% amino acid identity with the *Arabidopsis* GLP1 protein and lower identity (63%) with the GLP3 proteins. The *Pharbitis nil* GLP

shares 55 to 58% amino acid identity with all of the GLP subfamily 3 proteins.

By aligning 19 published germins and GLP cDNAs, we identified 16 completely conserved amino acid residues among all 19 clones from the seven species (data not shown). A further 11 amino acids are conserved in 18 of the 19 germin and GLP cDNAs. Among only the *Arabidopsis* GLPs, there are 31 identical amino acid residues. These residues are polydisperse throughout the proteins.

The most highly conserved region occurs near the middle of the proteins. From amino acids 107 to 119, there is a region with nearly 50% conserved identity. This conserved sequence is GxxP^h/pH^T/_hHP^G/_RA^T/_SE, where h signifies a hydrophobic residue. This is the same sequence that identifies a region of similarity with *Physarum* spherulin proteins [6, 36]. The conservation in this region is also variable among the various GLP subfamilies. This conservation is greatest in GLP subfamily 3. These GLPs have a conserved hydrophobic residue (either Leu or Met) at position 5 and a Gly at position 10, while all other germins and GLPs conserve a Pro at position 5 and an Arg at position 10. In the GLP subfamily 3, a leucine is present at position 5 of this sequence in GLP3a, GLP3b and SaGLP proteins and the methionine is present in this position in GLP1, BnGLP and PnGLP proteins. Whether this region of highest conserved identity among the GLPs bears any significance to enzyme function is not clear. However, the finding of class specificity in sequence conservation among the different family groups may indicate that the different family groups could play different roles within plants.

A pair of conserved cysteines is located near the N-terminus throughout the GLP family and may form a disulfide bond. No other conserved cysteines were identified. However, GLP3a, GLP3b, GLP4, GLP5

Table 3b. Differences between At-GLP3a and At-GLP3b.

	Position	Type	GLP3a	GLP3b	Mutation
1	120	point	T	C	Asp ²⁹ (silent)
2	146	point	A	T	Gly ³⁸ (silent)
3	209	point	A	C	Lys ⁵⁹ → Thr ⁵⁹
4	375	point	T	C	Val ¹¹⁴ (silent)
5	399	point	G	A	Gly ¹²² (silent)
6	411	point	T	C	Ser ¹²⁶ (silent)
7	414	point	T	C	Ala ¹²⁷ (silent)
8	429	point	A	G	Leu ¹³² (silent)
9	435	point	A	T	Thr ¹³⁴ (silent)
10	438	point	T	C	Leu ¹²⁵ (silent)
11	447	point	T	A	Gly ¹³⁸ (silent)
12	513	point	A	G	Leu ¹⁶⁰ (silent)
13	552	point	G	A	Gln ¹⁷³ (silent)
14	657	point	A	G	Gly ²⁰⁸ (silent)
15	675	point	A	T	3' UTR
16	693	deletion	T	–	3' UTR
17	701	deletion	C	–	3' UTR
18	756	deletion	–	TTCTTCTAATGATTGGT	3' UTR
19	802	deletion	G	–	3' UTR
20	805	deletion	–	T	3' UTR
21	807	deletion	T	–	3' UTR
22	813	deletion	C	–	3' UTR
23	834	deletion	–	C	3' UTR

and GLP7 also contain a single, unpaired cysteine. These unpaired cysteines are found in GLP4 near the N-terminus (C³), in GLP3a and GLP3b near the middle of the protein (C¹²⁰), in GLP5 at the C-terminus (C²⁰⁸) and in GLP7 at C⁵⁰.

Each of these germin-like proteins was analyzed for protein sorting signals and localization sites using the PSORT analysis tool, which identifies the existence of signal sequences using several methods [44, 48, 57]. With the exception of GLP4, which is predicted to contain a non-cleavable N-terminal sequence, all of the *Arabidopsis* GLPs contain a signal peptide between 17 and 23 amino acids in length. The GLP4 sequence lacks all elements of a cleavable N-terminal sequence, including: the charged residues near the N-terminus, the hydrophobic core and an alanine within the first 25 amino acids. The remaining signal sequences are presented in Table 4. Most of these sequences have charged residues in either the second or third position followed by a long stretch of hydrophobic amino acids. All signal sequences with the exception of GLP6 end with an alanine residue.

The secondary structure of each GLP was analyzed using the Protein Predict tool at EMBL, Heidelberg. The α -helical content is relatively low in all of these proteins, ranging from 3% to 20%. In contrast, the β -sheet ranges from 27% to 43%. The percent nonstructured (Loop) is ca. 55% for all GLPs. The location of these secondary structural elements along the GLP polypeptide backbones is shown in Figure 2.

As expected for all GLPs in which there is a predicted cleavable N-terminal signal sequence, each signal sequence contained an α -helix. In each of these cases, the α -helix extends until near the site of signal sequence cleavage. The GLP4 protein does not contain a cleavable N-terminal sequence and likewise does not contain an α -helical region near the N-terminus. Instead, this protein contains a predicted β -sheet structure near the N-terminus which is not found in the other proteins. In addition to the α -helix located at the N-terminus in most GLPs, there is another α -helix located at the C-terminus for 11 of the 12 GLPs. GLP5 does not contain this C-terminal α -helix.

Most of the predicted secondary structures found in the GLP gene family consist of 7 highly con-

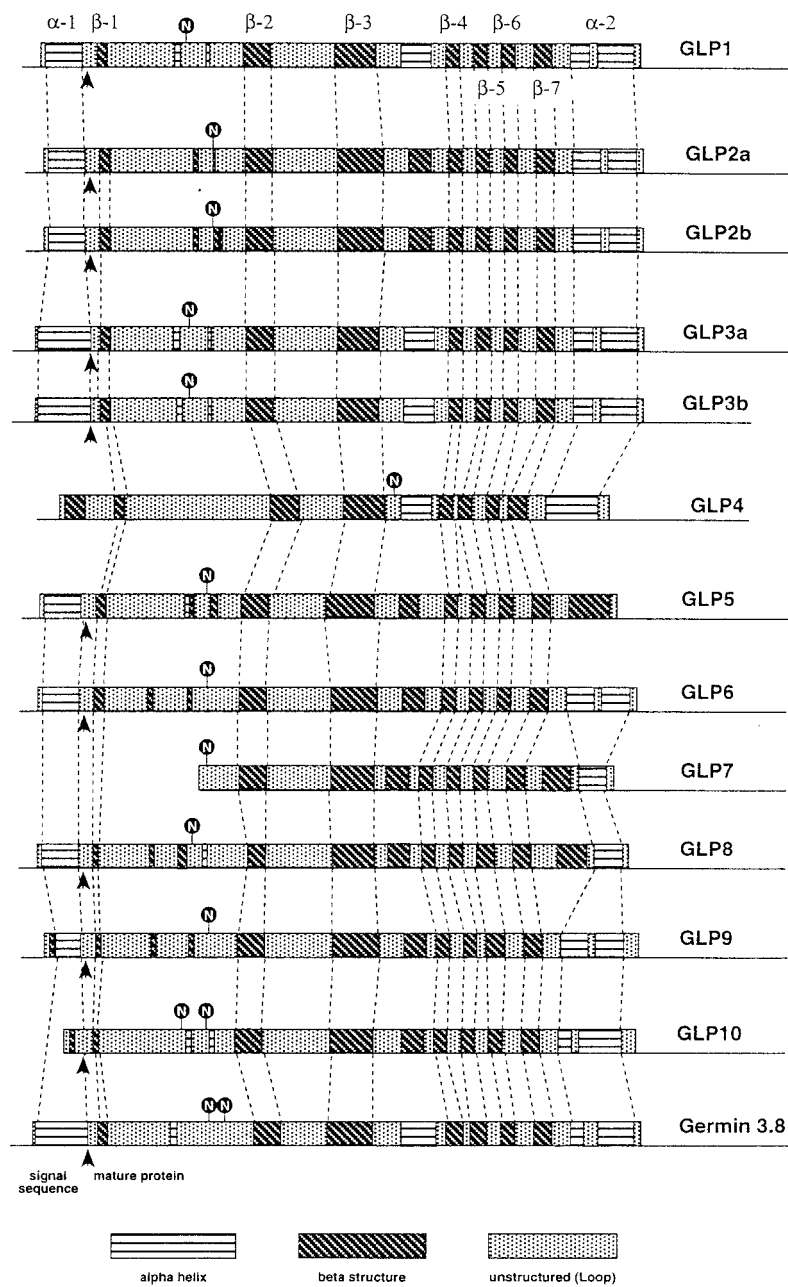


Figure 2. Predicted secondary structures of the GLP cDNAs. The protein sequences determined from each GLP cDNA was analyzed for folding patterns by the Protein Predict tool at EMBL in Heidelberg using the proteins predicted from the sequences listed in the legend to Figure 1. The identity of each GLP cDNA is presented at the right of each structure. Two conserved α -helices are labeled $\alpha-1$ and $\alpha-2$. The conserved β -structures are labeled $\beta-1$ through $\beta-7$. Predicted sites of N-glycosylation are indicated.

Table 4. Comparison of GLP proteins.

Protein	Pre-protein		Mature protein ^a		Signal sequence ^a	Potential N-glycosyl sites
	MW ^b	pI ^b	MW	pI		
GLP1	21559	9.3	19654	9.0	MLRTIFLLSL LFALSNA	59 (NTT)
GLP2a	23925	6.5	21567	6.2	MRVSQLVPF AIALVLSFV NA	77 (NVT)
GLP2b	23868	5.9	21510	5.6	MRVSQLVPF AIALVLSFV NA	77 (NVT)
GLP3a	21863	6.8	19564	6.3	MKMIIQIFFI ISLISTISFA	62 (NTS)
GLP3b	21836	6.3	19537	5.8	MKMIIQIFFI ISLISTISFA	62 (NTS)
GLP4	^c	–	21457	6.3	non-cleavable signal sequence	120 (NNT)
GLP5	21793	5.8	19397	5.8	MASPTLTL LLTTVSFFIS SSA	70 (NNT)
GLP6	24093	5.9	22167	5.3	MRVSKSLILI TLSALVIS	78 (NVT)
GLP7 ^d	–	–	–	–	–	5 (NVT)
GLP8	23033	8.9	20684	8.9	MARSMIPIFV TFNLVA AHMA LA	70 (NSS)
GLP9 ^e	23244	5.8	21071	5.8	<u>MTIK</u> SLSFLAALS LFALTLPVI A	77 (NVT)
GLP10 ^f	–	–	21733	7.9	...VIVLA	35 (NIT) 53 (NNT)

^aThe N-terminal signal sequences were identified using the PSORT protein sorting predictor tool at 'http://psort.nibb.ac.jp/'. The sequence of the mature protein was obtained by removing the predicted N-terminal signal sequences from the translated cDNA sequences.

^bThe molecular weight and pI analysis were performed using the ExPASy Protein Mw/pI tool at 'http://expasy.chuge.ch/ch2d/pi_tool.html'.

^cThe GLP4 protein contains a non-cleavable N-terminal signal sequence, therefore, this protein does not contain a 'pre' form.

^dThe GLP7 protein is missing ca. 70 N-terminal amino acids, and was hence not analyzed.

^eGLP9 is a nearly complete cDNA, but it is missing several amino acids from the N-terminus. The sequence provided here was determined from the GLP9 genomic sequence.

^fGLP10 is incomplete and could not be analyzed in a 'pre' form. The signal sequence and the N-terminus of the mature protein was evaluated by adding amino acids 1 to 15 from the GLP3 protein to the GLP 10 N-terminus and processing as with the others.

served β -pleated sheet domains, found primarily in the C-terminal half of the proteins. The N-terminal half of the proteins is less structured. Interestingly, most of these conserved β -structures are flanked by absolutely conserved amino acid residues. These conserved residues are frequently glycine. Glycine residues are often conserved in enzymes for flexibility in enzyme structure and function [59]. In addition to the major conserved β -structures, several smaller regions of secondary structure were also identified. Several α -helix and β -sheet regions were found occurring primarily in the large unstructured region between β -1 and β -2.

Finally, the structural elements were also determined for the wheat germin 3.8. As is shown in the figure, the wheat germin 3.8 protein is extremely similar to the *Arabidopsis* GLPs. It contains the same predicted α -helices and β -structures that are conserved throughout the *Arabidopsis* GLP families. These structural similarities confirm the relationship between the germains and the other GLPs. While it is

not clear that the enzymatic activities are conserved between the different subfamilies, the folding patterns of the proteins are probably very similar.

Possible sites of N-glycosylation are also illustrated in Figure 2. Ten of the 12 GLPs contain a single N-glycosylation site located between amino acids 60 and 80 (see Table 3). The GLP4 protein has lost this site of N-glycosylation, however, this protein has acquired a new site located at Asn-120. The GLP10 protein has acquired a second potential N-glycosylation site near by the conserved site. Thus, all *Arabidopsis* GLPs contain one or two putative sites of N-glycosylation as does wheat germin [29].

The molecular mass and the calculated pI for each GLP pre-protein and each mature GLP is also presented in Table 4. The molecular mass of each of the mature GLPs is 21 to 22 kDa. This is similar to the sizes of other germains and GLPs.

The GLP3b gene is physically linked to the polyubiquitin 4 gene

BLASTn homology searches [2] were performed on each of the GLP cDNAs. The BLAST search for one of these, GLP3b, identified a region of identity with the *Arabidopsis thaliana* polyubiquitin 4 gene (*ubq4*; GenBank accession number U33014 [12]). The GLP3b clone shares near 100% identity with the upstream region of the *ubq4* gene. The 3' end of the GLP3b transcript maps to only 500 nucleotides from the start of the *ubq4* gene. Thus, the genomic arrangement of the GLP3b-*ubq4* locus is presented in Figure 3. Like the wheat germins [36], the GLP3b gene does not contain introns.

The wheat gf-2.8 germin gene contains auxin regulatory elements in the proximal promoter [36] and is also responsive to IAA [27, 34]. Therefore, we searched the GLP3b promoter for the presence of putative regulatory elements. Two sequences similar to the SAUR A box reported by McClure *et al.* [42], (TGATAAAGG and TGACAAAA) were located at -357 to -349 and -204 to -196 upstream from the 5' end of the GLP3b clone respectively. In addition, a single putative auxin-responsive element, GTACCATGC, similar to the SAUR B' box sequences reported by McClure *et al.* [42], was located at -599 to -591 upstream from the 5' end of the GLP3b clone. Thus, the GLP3b gene may also be auxin-responsive. A TATA box, TCTCTATATAAAC, which is nearly identical (10/13) to the canonical plant TATA sequence [30], was identified at 36 nucleotides upstream from the 5' end of the GLP3b cDNA.

GLP9 resides on chromosome 4

Blast searches also revealed that the GLP9 gene is localized to a fragment of chromosome 4 (ESSA I contig fragment 1) that was recently deposited in the GenBank [7]. The GLP9 gene shown in Figure 3 is located between 145 kb and 155 kb of the ESSA I contig fragment 1.

The GLP9 cDNA is incomplete at its 5' end. Comparisons of the cDNA and genomic sequences permits us to identify those amino acids missing at the N-terminus of the preGLP9 protein. Analysis of these sequences indicates that the GLP9 reading frame ends 66 nt upstream from the 5' end of the GLP9 cDNA and identifies only a single methionine codon in frame with the GLP9 cDNA coding sequence. This in frame methionine codon is 11 nucleotides upstream from the

5' end of the GLP9 cDNA and encodes a Met-Thr-Ile-Lys tetrapeptide that we have taken as the true N-terminus of the preGLP9 protein. This sequence was used in all analyses of the preGLP9 protein (above).

Because the GLP9 gene is incomplete at the 5' end, the length of the 5' UTR is also unknown. Two potential TATA boxes were identified in the proximal GLP9 promoter. Both of these sequences have similar degrees of identity (10/13) with the canonical plant TATA box [30]. The first is located 77 nucleotides upstream from the methionine start codon and the second is located 100 nucleotides upstream from the methionine start codon.

The 2471 bp intergenic region between the GLP9 coding region and the coding region of the 5'-flanking genes, a rat SVP homologue must contain promoter sequences required to drive each of these genes. Analysis of this intergenic region demonstrated no significant stretches of either direct or inverted repeats. Like the GLP3b promoter, an auxin A regulatory element, TAATGAAAG [42], was identified in the GLP9 promoter at -280 to -272 upstream from the methionine start codon. In addition, two Box 1 light-regulating elements [24] TTTCAA were identified in the proximal GLP9 promoter at -295 to -289 and -182 to -175 upstream from the methionine start codon. These sequences may function to express the GLP9 protein in a circadian oscillation as has been found with other GLPs (26). In contrast to the GLP3b gene and the true germin genes that have been previously analyzed, the GLP9 gene contains a single 92 bp intron interrupting the coding region of the preGLP9 protein at amino acid Val-44.

Discussion

Our interest in germin-like proteins stems from the recent observations that germin-like proteins may function in plant defenses [19, 28, 60]. The *Arabidopsis* GLPs are not yet known to be associated with an enzyme activity that can be directly linked to plant defenses. Nevertheless, the similarity of the *Arabidopsis* GLP proteins at the amino acid sequence and protein structural levels to the cereal germin proteins indicates that these proteins are highly related and may share biological activities. To examine the role of germin-like proteins in plant defenses we have begun by characterizing all of the germin-like cDNAs that have been isolated from numerous *Arabidopsis* cDNA libraries. This analysis has demonstrated that there are

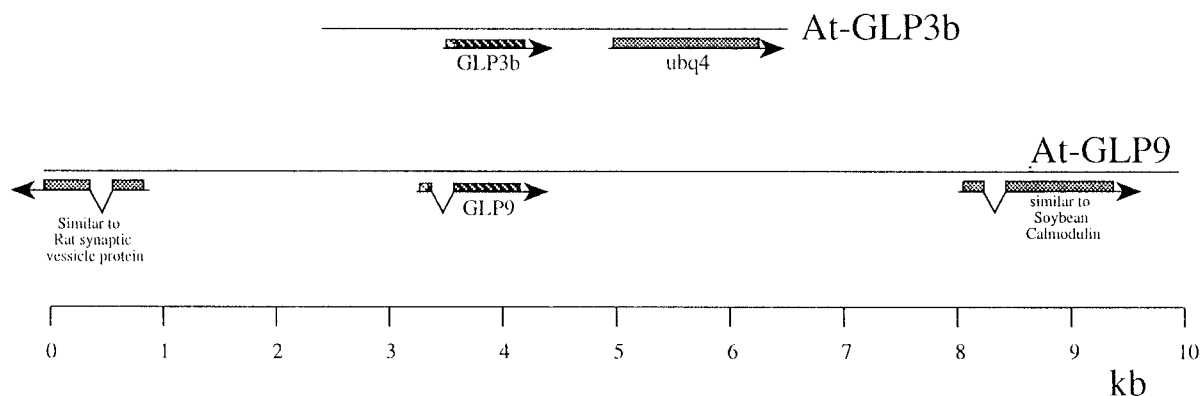


Figure 3. Structure of genes encoding GLP3b and GLP9. DNA sequences for the *Arabidopsis thaliana* GLP3b and GLP9 genes were identified in GenBank searches. The genomic sequences are represented by the horizontal lines. Various identified genes are indicated by the hatched boxes located on the arrows. The direction of the arrows indicate the direction of transcription. The GLP genes are indicated by the striped boxes in the center of the fragment. Introns are presented as V-shaped structures below the various genes. A scale of gene size is presented below the chromosomal fragments.

at least 12 transcribed GLP genes within the *Arabidopsis* genome. The proteins encoded by these mRNAs all share a minimum of 31% amino acid identity. These 12 sequences fall into four related subfamilies which show a higher level of conserved amino acid identity among the subfamily members.

With a single exception, all of the translated proteins contain an N-terminal signal sequence. The N-terminal signal sequences vary in length from 17 to 23 amino acids and all show characteristics of classical N-terminal signal sequences, (i.e., an α -helical core regions flanked by charged amino acids, ending in an alanine). With the exception of GLP4 which does not contain a cleavable N-terminal signal sequence, the accumulation of the proteins are predicted to occur at three different sites; outside the cell, in the vacuole and at the plasma membrane.

This predicted pattern of intracellular localization is similar to that of the PR proteins in plants [40]. PR proteins are also found as multigene families having both acidic and basic forms that accumulate both intracellularly and extracellularly [45, 55]. However, it is not known whether the acidic and basic forms of the *Arabidopsis* GLPs accumulate at different cellular locations as the PR proteins do.

Dratewka-Kos *et al.* [18] indicate that germin contains a long stretch of uncharged amino acids that may mediate its homopentameric assembly. We therefore examined the GLP clones for distribution of charge. The average overall density of charged residues is 14.8% for all of the mature GLPs. However, the charge distribution is heterogeneous in the GLP family. There

are two regions with a low charge density having only 2.5% charged residues throughout these regions. The first region is located between amino acids 60 to 99 in the β -1 to β -2 loop and contains the conserved site of N-glycosylation. The second stretch is found at amino acids 169 to 198 and overlaps both the β -6 and β -7 regions. For comparison, the average charge density was calculated for the three remaining regions (N-terminus to amino acid 59; amino acids 100 to 168; and amino acids 199 to C-terminus). These three regions show 24.5%, 19.2% and 24.8% charged residues, respectively. The uncharged region in wheat germin is located at amino acids 147 to 180 of the mature protein [18]. This corresponds to our charge deficient region 2 that is present among the *Arabidopsis* GLPs. If this region is responsible for the homopentameric nature of these proteins as proposed by Dratewka-Kos *et al.*, it is possible that the conserved β -6 and β -7 structures may mediate this interaction.

The finding that germin-like proteins may be glycosylated is not surprising since many secreted proteins in both plants and animals are glycosylated. However, the finding of a conserved glycosylation site in all of the GLPs is intriguing. In 11 of the 12 distinct clones examined, there are conserved glycosylation sites within a stretch of only 10 amino acids. In addition, the site of N-glycosylation in wheat germin is at position 47 and/or 52 of the mature polypeptide [18]. Thus, this site of glycosylation within the germins and GLPs has been conserved in both monocots and dicots, over a timeframe of about 125 million years [17].

The mRNAs produced from several of the GLP genes were found to be quite heterogeneous at their respective 3' ends. Correct mRNA 3' formation is an essential step in gene expression [58], and consensus sequences have been identified for a variety of eukaryotic mRNAs including plants [15]. For the majority of the GLP mRNAs, apparent polyadenylation sites similar to those identified by Joshi [31] were recognized. However, when the polyadenylation sites of 10 of the GLP1 mRNAs were examined, 50% of the mRNAs did not contain a consensus polyadenylation signal in the proximity of the polyadenylation sites. In those mRNAs that did not utilize a consensus polyadenylation signal, there was a sequence, CATTG, located 15 to 19 nucleotides upstream from the polyadenylation site that has been previously implicated in polyadenylation [4]. This signal is complementary to regions within the small nuclear RNA U4, and suggests that U4 small nuclear ribonucleoproteins (snRNPs) may mediate polyadenylation in a manner similar to the role of U1 snRNPs in splicing [5]. The finding that this signal is as efficient as the ATTAAG signal indicates that mRNA processing and polyadenylation is heterogeneous in plants. Further evidence of this comes from the GLP5 mRNAs in which non-consensus polyadenylation signals are used 75% of the time.

The proximity of the end of the GLP3b transcript to the *ubq4* start signals should affect the expression of one or both of these genes. It should be noted though that both of these genes are apparently functional because multiple GLP3b transcripts as well as multiple *ubq4* transcripts have been cloned as ESTs. However, when transgenic plants were prepared expressing GUS under the control of the *ubq4* promoter, this promoter was identified as a 'poor' responder (J. Callis, personal communication). It may be that this observed poor level of expression of the *ubq4* promoter in transgenic plants can be attributed to a functional gene in the proximal promoter. The proximity of the GLP3b gene to the *ubq4* gene raises some interesting questions regarding the genome organization and structure in *Arabidopsis*. We have found such close proximity of one gene to another is not uncommon. In separate studies we have identified an EST located within the promoter of the *A. thaliana* UMP synthase gene (Kafer and Thornburg, unpublished). The 3' end of this EST maps to within 520 nucleotides of the 5' end of the UMP synthase cDNA. Thus, intergenic regions within the *Arabidopsis* genome are frequently quite short.

Whether additional GLP sequences will be found in association with regions further upstream from

the polyubiquitin genes will require further analysis. However, it should be noted that the polyubiquitin gene family and the germin-like protein gene family are both large, relatively homologous gene families. It may be that additional associations between the two gene families will be noted with time.

Acknowledgements

This work was sponsored by funds from the Carver Trust, the Hatch Act, and State of Iowa funds. The authors would like to thank Dr François Bernier for sharing the results of his studies prior to their publication.

References

1. ABRC: Library information for ESTs. In: *Arabidopsis Biological Resource Center: Seed and DNA Stock List*, p. 239. Ohio State University, 1735 Neil Avenue, 309 Botany & Zoology Bldg., Columbus, OH 43210, USA (1995).
2. Altschul SF, Fish W, Miller W, Myers EW, Lipman DJ: Basic local alignment search tool. *J Mol Biol* 215: 403–410 (1990).
3. Apostol I, Heinsteins PF, Low PS: Rapid stimulation of an oxidative burst during elicitation of cultured plant cells. *Plant Physiol* 90: 109–116 (1989).
4. Benoit C, O'Hare K, Breathnach R, Chambon P: The ovalbumin gene: sequence of putative control regions. *Nucl Acids Res* 8: 127–142 (1980).
5. Berget SM: Are U4 small nuclear ribonucleoproteins involved in polyadenylation? *Nature* 309: 179–182 (1984).
6. Bernier F, Lemieux G, Pallotta C: Gene families encode the major encystment-specific proteins of *Physarum polycephalum* plasmodia. *Gene* 59: 265–277 (1987).
7. Bevan M, Stiekema W, Murphy G, Wambutt R, Pohl T, Terry N, Kreis M, Kavanagh T, Entian KD, Rieger M, James R, Puigdomenech P, Hatzopoulos P, Obermaier B, Duesterhoft A, Jones J, Palme K, Ansoerge W, Delseny M, Bancroft I, Mewes HW, Schueller C, Chalwatzis N: *Arabidopsis thaliana* DNA chromosome 4, ESSA I contig fragment No. 1. GenBank accession number Z97336 (1997).
8. Bjellqvist B, Hughes GJ, Pasquali C, Paquet N, Ravier F, Sanchez J-C, Frutiger S, Hochstrasser DF: The focusing positions of polypeptides in immobilized pH gradients can be predicted from their amino acid sequences. *Electrophoresis* 14: 1023–1031 (1993).
9. Bradley DJ, Kjellbom P, Lamb CJ: Elicitor and wound-induced oxidative cross-linking of a proline-rich plant cell wall protein: a novel, rapid defense response. *Cell* 70: 21–30 (1992).
10. Brisson LF, Tenhaken R, Lamb C: Function of oxidative cross-linking of cell wall structural proteins in plant disease resistance. *Plant Cell* 6: 1703–1712 (1994).
11. Callis J: *Arabidopsis thaliana* polyubiquitin (*ubq4*) gene. GenBank accession number U33014 (1995).
12. Callis J, Carpenter T, Sun CW, Vierstra RD: Structure and evolution of genes encoding polyubiquitin and ubiquitin-like

- proteins in *Arabidopsis thaliana* ecotype Columbia. *Genetics* 139: 921–939 (1995).
13. Chen Z, Silva H, Kelessig DF: Active oxygen species in the induction of plant systemic acquired resistance by salicylic acid. *Science* 262: 1883–1886 (1994).
 14. Chen Z, Malamy J, Henning J, Conrath U, Sanchez-Casas P, Silva H, Ricigliano J, Klessig DF: Induction, modification and transduction of the salicylic acid signal in plant defense responses. *Proc Natl Acad Sci USA* 92: 4134–4137 (1995).
 15. Dean C, Tamaki S, Dunsmuir P, Favreau M, Katayama C, Dooner H, Bedbrook J: mRNA transcripts of several plant genes are polyadenylated at multiple sites *in vivo*. *Nucl Acids Res* 14: 2229–2240 (1986).
 16. Doman J-M, Dumas B, Lainé E, Meyer Y, Alain D, David H: Three glycosylated polypeptides secreted by several embryonic cell cultures of pine show highly specific serological affinity to antibodies directed against the wheat germin apoprotein monomer. *Plant Physiol* 108: 141–148 (1995).
 17. Doyle JA: Origin of angiosperms. *Annu Rev Ecol Syst* 9: 365–392 (1978).
 18. Dratewka-Kos E, Rahman S, Grzelczak Z, Kennedy TD, Murray RK, Lane BG: Polypeptide structure of germin as deduced from cDNA sequencing. *J Biol Chem* 264: 4896–4900 (1989).
 19. Dumas B, Sailland A, Cheviet JP, Freyssinet G, Pallett K: Identification of barley oxalate oxidase as a germin-like protein. *C.R. Acad Sci III* 316: 793–798 (1993).
 20. Dumas B, Freyssinet G, Pallett K: Tissue-specific expression of germin-like oxalate oxidase during development and fungal infection of barley seedlings. *Plant Physiol* 107: 1091–1096 (1995).
 21. Fry SC: Cross-linking of matrix polymers in the growing cell walls of angiosperms. *Annu Rev Plant Physiol* 37: 165–186 (1986).
 22. GCG (Version 9.0) Wisconsin Package, Genetics Computer Group, Madison WI.
 23. Graham RA, Thornburg RW: DNA sequence of UDP Glucose:Indole-3-acetate β -D-glucosyl transferase from *Arabidopsis thaliana* (U81293). *Plant Physiol* 113: 1004 (1997).
 24. Green PJ, Kay SA, Chua N-H: Sequence-specific interactions of a pea nuclear factor with light responsive elements upstream of the *rbcS-3A* gene. *EMBO J* 6: 2543–2549 (1987).
 25. Grellet F, Cooke R, Laudie M, Raynal M, Delseny M: A *thaliana* mRNA for germin type 2 protein. GenBank accession number X91957 (1995).
 26. Heintzen C, Fischer R, Melzer S, Kappeler S, Apel K, Staiger D: 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*. *Plant Physiol* 106: 905–615 (1994).
 27. Hurkman WJ, Tanaka CK: Effect of salt stress on germin gene expression in barely roots. *Plant Physiol* 110: 971–977 (1996).
 28. Hurkman WJ, Tanaka CK: Germin gene expression is induced in wheat leaves by powdery mildew infection. *Plant Physiol* 111: 735–739 (1996).
 29. Jaikaran ASI, Kennedy TD, Dratewka-Kos E, Lane BG: Covalently bonded and adventitious glycans in germin. *J Biol Chem* 265: 12503–12512 (1990).
 30. Joshi CP: An inspection of the domain between putative TATA box and translation start site in 79 plant genes. *Nucl Acids Res* 15: 6643–6653 (1987).
 31. Joshi PC: Putative polyadenylation signals in nuclear genes of higher plants: A compilation and analysis. *Nucl Acids Res* 15: 9627–9640 (1987).
 32. Kozak M: Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes. *Cell* 44: 283–292 (1986).
 33. Lachman PJ: A common form of killing. *Nature* 321: 560 (1986).
 34. Lane BG: Oxalate, germin and the extracellular matrix of higher plants. *FASEB J* 7: 294–301 (1994).
 35. Lane BG, Grzelczak ZF, Kennedy TD, Kajjola R, Orr J, D'Agostino S, Jaikaran A: Germin: compartmentation of two forms of the protein by washing growing wheat embryos. *Biochem. Cell Biol* 64: 1025–1037 (1986).
 36. Lane BG, Bernier F, Dratewka-Kos E, Shafai R, Kennedy TD, Caron P, Munro JR, Vaughan T, Walters D, Altomare F: Homologies between members of the germin gene family in hexaploid wheat and similarities between these wheat germins and certain *Physarum* spherulins. *J Biol Chem* 266: 10461–10469 (1991).
 37. Lane BG, Cuming AC, Frégeau J, Carpita NC, Hurkman WJ, Bernier F, Dratewka-Kos E, Kennedy TD: Germin isoforms are discrete temporal markers of wheat development. Pseudogermin is a uniquely thermostable water-soluble oligomeric protein in ungerminated embryos and like germin in germinated embryos, it is incorporated into cell walls. *Eur J Biochem* 209: 961–969 (1992).
 38. Lane BG, Dunwell JM, Ray JA, Schmitt MR, Cuming AC: Germin, a protein marker of early plant development is an oxalate oxidase. *J Biol Chem* 268: 12239–12242 (1993).
 39. Levine A, Tenhaken R, Dixon R, Lamb C: H₂O₂ from the oxidative burst orchestrates the plant hypersensitive disease resistance response. *Cell* 79: 583–593 (1994).
 40. Linthorst HJM, van Loon LC, van Rossum CMA, Mayer A, Bol JF, van Roekel JSC, Meulenhoff EJS, Josien S, Cornelissen BJC: Analysis of acidic and basic chitinases from tobacco and petunia and their constitutive expression in transgenic tobacco. *Mol Plant-Microbe Interact* 3: 252–258 (1990).
 41. Marck C: 'DNA Strider': a 'C' program for the fast analysis of DNA and protein sequences on the Apple Macintosh family of computers. *Nucl Acids Res* 16: 1829–1836 (1988).
 42. McClure BA, Guilfoyle TJ: Characterization of a class of small auxin-inducible polyadenylated RNAs. *Plant Mol Biol* 9: 611–623 (1987).
 43. McCubbin WC, Cyril MK, Kennedy TD, Lane BG: Germin: physicochemical properties of the glycoprotein which signals the onset of growth in the germinating wheat embryo. *Biochem Cell Biol* 65: 1039–1048 (1987).
 44. McGeoch DJ: On the predictive recognition of signal peptide sequences. *Virus Res* 3: 271 (1985).
 45. Melchers LS, Sela-Buurlage MB, Vloemans SA, Woloshuk CP, van Roekel JSC, Pen J, van den Elzen PJM, Cornelissen BJC: Extracellular targeting of the vacuolar tobacco proteins AP24, chitinase, and β -1,3-glucanase in transgenic plants. *Plant Mol Biol* 21: 583–593 (1993).
 46. Membré N, Berna A, Neutelings G, David A, David H, Staiger D, Vásquez JS, Raynal M, Delseny M, Bernier F: cDNA sequence, genomic organization and differential expression of three *Arabidopsis* genes for germin/oxalate oxidase-like proteins. *Plant Mol Biol* 35: 459–469 (1997).
 47. Michalowski CB, Bohnert HJ: Nucleotide sequence of a root specific transcript encoding a germin-like protein from the halophyte *Mesembryanthemum crystallinum*. *Plant Physiol* 100: 537–538 (1992).
 48. Nakai K, Kanehisa M: Expert system for predicting protein localization sites in Gram-negative bacteria. *Proteins Struct Funct Genet* 11: 95–110 (1991).

49. Nakai K, Kanehisa M: A knowledge base for predicting protein localization sites in eukaryotic cells. *Genomics* 14: 897–911 (1992).
50. Ono M, Sage-Ono K, Kamada H, Harada H: A cDNA encoding a germin-like protein from *Pharbitis nil*. GenBank accession number D45425 (1995).
51. Raynal, M: *A. thaliana* mRNA for germin 1 protein. GenBank accession number: X91921 (1995).
52. Rost B, Sander C: Prediction of protein structure at better than 70% accuracy. *J Mol Biol* 232: 584–599 (1993).
53. Rost B, Sander C, Schneider R: PHD: an automatic mail server for protein secondary structure prediction. *CABIOS* 10: 53–60 (1994).
54. Sun C-W, Callis J: Recent stable insertion of mitochondrial DNA into an *Arabidopsis* polyubiquitin gene by nonhomologous recombination. *Plant Cell* 5: 97–107 (1993).
55. van den Bulcke M, Bauw G, Castresana C, van Montagu M, Vandekerckhove J: Characterization of vacuolar and extra-cellular β -(1,3)-glucanases of tobacco: Evidence for a strictly compartmentalized plant defense system. *Proc Natl Acad Sci USA* 86: 2673–2677 (1989).
56. Varner JE, Lin LS: Plant cell wall architecture. *Cell* 56: 231–239 (1989).
57. von Heijne G: A new method for predicting signal sequence cleavage sites. *Nucl Acids Res* 14: 4683–4690 (1986).
58. Wickens M: How the messenger got its tail: addition of poly(A) in the nucleus. *Trends Biochem Sci* 15: 277–281 (1991).
59. Yan BX, Sun YQ: Glycine residues provide flexibility for enzyme active sites. *J Biol Chem* 272: 3190–3194 (1997).
60. Zhang Z, Collinge DB, Thordal-Christensen H: Germin-like oxalate oxidase, a H_2O_2 -producing enzyme, accumulates in barley attacked by the powdery mildew fungus. *Plant J* 8: 139–145 (1995).