Available online at www.sciencedirect.com
SCIENCE DIRECT®
Phytochemistry 65 (2004) 43–58
PHYTOCHEMISTRY
www.elsevier.com/locate/phytochem

The S8 serine, C1A cysteine and A1 aspartic protease families in Arabidopsis

Eric P. Beers[a,*], Alan M. Jones[b], Allan W. Dickerman[c]

[a]Department of Horticulture, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061, USA
[b]Department of Biology, University of North Carolina, Chapel Hill, NC 27599, USA
[c]Virginia Bioinformatics Institute, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061, USA

Received 4 June 2003; received in revised form 25 August 2003

Abstract

The *Arabidopsis thaliana* genome has over 550 protease sequences representing all five catalytic types: serine, cysteine, aspartic acid, metallo and threonine (MEROPS peptidase database, http://merops.sanger.ac.uk/), which probably reflect a wide variety of as yet unidentified functions performed by plant proteases. Recent indications that the 26S proteasome, a T1 family-threonine protease, is a regulator of light and hormone responsive signal transduction highlight the potential of proteases to participate in many aspects of plant growth and development. Recent discoveries that proteases are required for stomatal distribution, embryo development and disease resistance point to wider roles for four additional multigene families that include some of the most frequently studied (yet poorly understood) plant proteases: the subtilisin-like, serine proteases (family S8), the papain-like, cysteine proteases (family C1A), the pepsin-like, aspartic proteases (family A1) and the plant matrixin, metalloproteases (family M10A). In this report, 54 subtilisin-like, 30 papain-like and 59 pepsin-like proteases from Arabidopsis, are compared with S8, C1A and A1 proteases known from other plant species at the functional, phylogenetic and gene structure levels. Examples of structural conservation between S8, C1A and A1 genes from rice, barley, tomato and soybean and those from Arabidopsis are noted, indicating that some common, essential plant protease roles were established before the divergence of monocots and eudicots. Numerous examples of tandem duplications of protease genes and evidence for a variety of restricted expression patterns suggest that a high degree of specialization exists among proteases within each family. We propose that comprehensive analysis of the functions of these genes in Arabidopsis will firmly establish serine, cysteine and aspartic proteases as regulators and effectors of a wide range of plant processes.

Keywords: *Arabidopsis thaliana*; Crucifereae; Papain; Subtilisin; Pepsin; Cysteine protease; Serine protease; Aspartic protease

1. Introduction

1.1. Proteases as regulators of plant growth and development

Complete proteolysis and site-specific limited proteolysis are complementary, selective mechanisms that act in concert with other post-translational modifications to control the half-lives, subcellular trafficking, and activities of proteins. Proteases are required for a broad range of genetically programmed and inducible processes beyond classical protease roles in starvation, stress response and nutrient remobilization. Targeted proteolysis mediated by the ubiquitin/26S proteasome pathway is important to most aspects of plant biology (reviewed by Vierstra, 2003), including, for example, gibberellin- (Fu et al., 2002) and auxin- (Estelle, 2001) mediated responses and light perception (Schwechheimer and Deng, 2001). Additionally, experiments with Arabidopsis and tomato have revealed new roles for subtilisin-like (S8 family), papain-like (C1A family) and pepsin-like (A1 family) proteases as mediators of plant development and disease resistance (protease family names are those used by the MEROPS peptidase database, http://merops.sanger.ac.uk/). These results for S8, C1A and A1 proteases, as well as recent discoveries that some serine carboxypeptidase-like proteins (S10 family) function as acyltransferases in secondary metabolism rather than as hydrolases (Lehfeldt

* Corresponding author. Tel.: +1-540-231-3210; fax: 1-540-231-3083.
E-mail address: ebeers@vt.edu (E.P. Beers).

0031-9422/$ - see front matter © 2003 Elsevier Ltd. All rights reserved.
doi:10.1016/j.phytochem.2003.09.005

et al., 2000; Li and Steffens, 2000), indicate that a comprehensive functional analysis of predicted proteases would yield novel insights regarding a variety of plant processes.

1.2. Structural features of plant S8, C1A and A1 family proteases

The S8, C1A, and A1 protease families described in this report are among the largest multigene protease families known in Arabidopsis, with 54, 30 and 59 genes, respectively. Well-known representatives from these protease families are synthesized as preproproteins that undergo proteolytic processing of the pre and pro domains to yield the mature, active protease. The structure of the best-known plant subtilisin-like protease, preprocucumisin (EC 3.4.21.25) (Yamagata et al., 1994), is diagrammed in Fig. 1 along with structures of prepropapain (EC 3.4.22.2) (Groves et al., 1996) and preprophytepsin (EC 3.4.23.40) (Glathe et al., 1998). The subtilisin-like proteases exhibit a catalytic triad consisting of Asp30, His94 and Ser415 (mature cucumisin numbering, Yamagata et al., 1994) and the protease-associated (PA) domain (Fig. 1A). The PA domain is thought to mediate protein–protein interaction between protease and substrate (Mahon and Bateman, 2000). C1A proteases exhibit the characteristic papain protease unit consisting of the catalytic dyad, Cys25 and His159 (using papain numbering, Groves et al., 1996), and an Asn175 residue important for proper orientation of the His side chain (Fig. 1B). The C1A N-terminal prodomains are regulators of targeting (Holwerda et al., 1992; Ahmed et al., 2000), folding, and activity (Mach et al., 1994; Tao et al., 1994; Taylor et al., 1995). Active sites for the representative A1 family, pepsin-like protease (Fig. 1C) are centered at Asp36 and Asp223 (barley phytepsin numbering, Kervinen et al., 1999). The core of each active site is the Asp-Thr/Ser-Gly (DT/SG) motif.

The potential of the S8, C1A and A1 proteases as important regulatory proteins is only now becoming apparent. Fortunately, these protease families include members with long histories as subjects of structural, biochemical and molecular analyses, making them ideal subjects for efficient functional genomics projects. With the completion of the Arabidopsis genome comes the opportunity to explore the roles of S8, C1A and A1 enzymes from a single plant species. Doing so will address the question of functional redundancy and determine the significance of gene duplication, restricted expression and possible functional evolution of the catalytic mechanisms

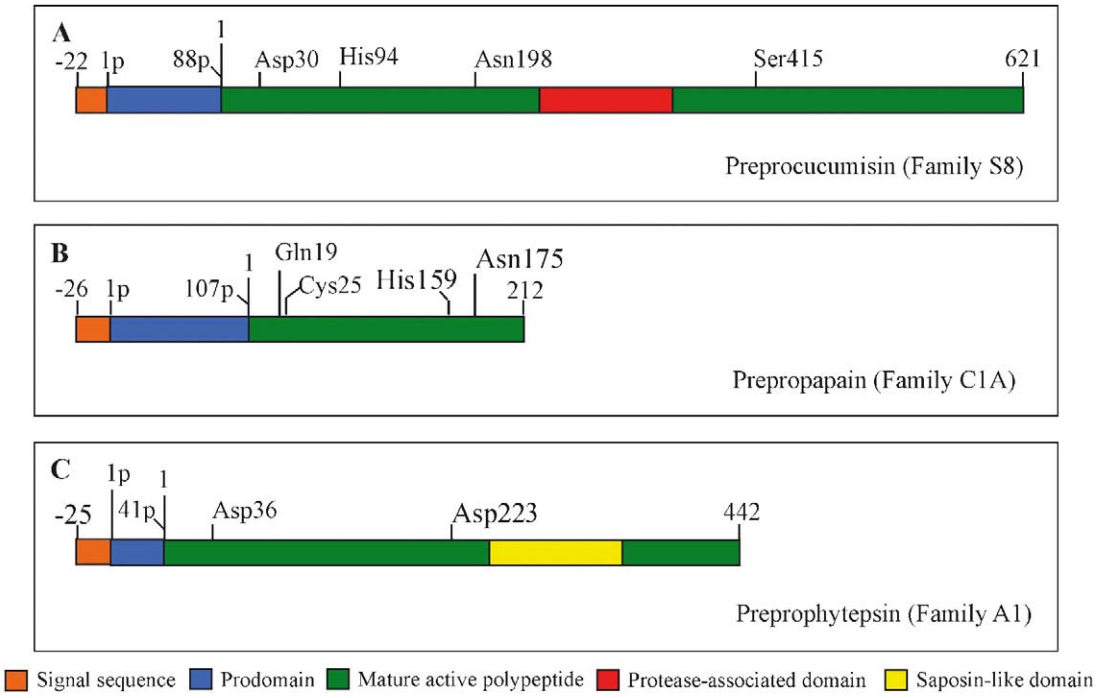


Fig. 1. Schematic representation of structures of preproproteins for the best known members of plant S8, C1A and A1 family proteases, preprocucumisin (A), prepropapain (B) and preprophytepsin (C). Signal sequences are indicated by negative numbers, and each prodomain is indicated using numbers followed by "p" . Numbering for the mature enzyme indicates positions of the N and C termini. For preprocucumisin (A) residues of the catalytic triad, Asp30, His94 and Ser415, are labeled, as is a conserved Asn (Asn198) involved in oxyanion hole stabilization. The core conserved region of the protease-associated (PA) domain, predicted to be involved in protease-substrate interaction (Mahon and Bateman, 2000), is also indicated. For prepropapain (B) the active site residues Cys25, His159 and Asn175 are shown along with the conserved Gln19 residue. For preprophytepsin (C) the two active site residues Asp36 and Asp223 are labeled and the position of the plant-specific, saposin-like domain (Kervinen et al., 1999) is indicated. While only a minority of the predicted plant A1 family proteases contain the saposin-like domain found in preprophytepsin, the barley enzyme shown here was chosen because it is the most extensively characterized representative of the plant pepsin-like enzymes.

of closely related members of multigene families of proteases as factors contributing to functional specialization. Toward this end, an analysis of the phylogenies, intron/exon arrangements, chromosome positions, estimated relative transcript levels and availability of insertional mutants for annotated Arabidopsis S8, C1A and A1 family genes is presented here as a tool for genetic studies.

2. S8, C1A and A1 proteases: functional studies, gene expression, phylogenetic analyses and gene structure comparisons

2.1. S8 proteases

Sequences predicted to code for S8 family proteases are known from all kingdoms (see MEROPS). The 54 Arabidopsis subtilisins (Table 1), as well as those reported from other plant species, are most similar to the bacterial S8A subfamily of subtilisins. There is only one member (CAB45880) of the S8C subfamily (tripeptidyl-peptidase II) of proteases, and no members of the S8B subfamily (the proprotein convertases) in Arabidopsis. Exceptions to the subtilisin structure (Fig. 1) are noted in Table 1 and include BAB09207, a potential pseudogene lacking both the Asp30 active-site residue and any representation among reported ESTs, and the SKI-1-like protein, NP197467. Despite its homology to S8A proteases, NP197467 lacks the PA domain present in other Arabidopsis subtilisins. NP197467 most closely resembles the mammalian SKI-1 proprotein convertase that catalyzes proteolytic activation of the precursors to sterol-regulated element-binding proteins (Seidah et al., 1999; Elagoz et al., 2001). Although the S8 (and the C1A and A1) plant enzymes are generally thought of as secretory pathway proteins, iPSORT (http://hypothesiscreator.net/iPSORT/) analyses indicated that not all enzymes assembled for this report possess a predicted signal sequence. Proteins found to have no obvious signal sequence or predicted to localize to chloroplasts or mitochondria are present in all three protease families described here (Tables 1–3). Experimentally determined localizations have not been reported for the vast majority of S8, C1A and A1 plant proteases, however.

Four recent reports demonstrate the importance of two subtilisin-like protease genes, *SDD1* and *ALE1*, in stomatal density and distribution (Berger and Altmann, 2000; Von Groll et al., 2002; Schluter et al., 2003) and proper epidermis formation at the endosperm-embryo interface (Tanaka et al., 2001), respectively. Loss of *SDD1* function results in a 2.5-fold increase in stomatal density. Compared to wild type plants, *sdd1-1* plants exhibited elevated CO_2 fixation when plants grown under low light conditions were transferred to high light intensities (Schluter et al., 2003). The experiments with *sdd1-1* plants illustrate the potential for proteases other than the 26S proteasome to have significant impact on plant productivity via mechanisms apparently not associated with bulk proteolysis. It has been suggested that SDD1 and ALE1 act as proprotein convertases yielding as yet unidentified bioactive peptides (Berger and Altmann, 2000; Tanaka et al., 2001), although direct evidence for plant subtilisin-like proteases acting as regulatory proprotein convertases has not yet been presented.

The *SDD1* and *ALE1* genes for Arabidopsis subtilisin-like proteases are not the only plant S8 enzymes investigated within recent years. The tomato genome also includes multiple genes that encode subtilisin-like proteases (Meichtry et al., 1999). Pathogens induce expression of *P69A* (JC6119) in tomato (Tornero et al., 1996). *TMP* (AAF13299) (Riggs et al., 2001) and *LIM9* (BAA04839) (Kobayashi et al., 1994; Taylor et al., 1997) expression has been linked to microsporogenesis in tomato and lily, respectively; although antisense experiments indicate that full expression of *TMP* is not essential for microsporogenesis (Riggs et al., 2001). Actinorhizal nodule development in *Alnus* (Ribeiro et al., 1995) is associated with expression of *ag12* (S52769) and a seed coat-specific subtilisin (SCS1, CAB87246) has been reported for soybean (Batchelor et al., 2000). In Arabidopsis, *ARA12* is expressed in all organs with highest levels detected in developing siliques (Ribeiro et al., 1995). Also in Arabidopsis, expression of the S8 protease gene *AIR3* is linked to lateral root emergence (Neuteboom et al., 1999), *XSP1* is expressed in tracheary elements (Beers and Zhao, 2001) and *SLP2* and *SLP3* are induced by stress (Golldack et al., 2003). Estimates of S8 protease gene transcript levels determined from Massively Parallel Signature Sequencing (MPSS) (Brenner et al., 2000), from Arabidopsis cDNA libraries (http://dbixs001.dbi.udel.edu/MPSS4/java.html) revealed that 32 (60%) of the subtilisin genes are represented (Table 1). Of these, only four genes are strongly expressed (i.e., normalized transcript abundance is greater than 1000 parts per million). Two of the four strongly expressed subtilisins, *ARA12* and *SLP2* (Ribeiro et al., 1995; Golldack et al., 2003), have been described experimentally, as mentioned above. Very low, 1–10 ppm, levels or absence of transcripts in the MPSS data set does not necessarily indicate a lack of importance for individual protease genes, however. For example, normalized transcript abundance for the stomatal density and distribution subtilisin gene *SDD1* (Berger and Altmann, 2000) is only 18 ppm and transcripts for other genes with cell-type-specific roles may also be present at very low levels on an organ-wide basis. Despite the relatively large size of the subtilisin gene family in Arabidopsis, the family-wide total transcript value, 8974 ppm, is only 48 and 55% of the values determined for the C1A and A1 families, respectively. In addition to

Table 1
Arabidopsis subtilisin-like serine proteases (S8 family)

Protein ID[a]	AGI code[b]	Chrom, position	T-DNA[c]	MPSS[d] ppm	iPSORT[e]	Published name	Notes/references
S8-1							
BAB09626	At5g58810	V, 23461868	+	0	S		
BAB09627	At5g58820	V, 23465862	+	37	M		
BAB09628	At5g58830	V, 23469879	–	0/+	S		
BAB09629	At5g58840	V,23472949	+	156	S		
BAB10784	At5g59090	V, 23566031	+	196	M		
BAB10785	At5g59100	V, 23572857	+	0/+	S		
BAB09758	At5g59120	V, 23578803	+	270	S		
BAB09759	At5g59130	V, 23584098	+	9	NS		
CAB51180	At3g46840	III, 17259971	+	0	S		
CAB51181	At3g46850	III, 17265298	–	0	S		
AAB95271	At2g39850	II, 16579170	+	17	S		
CAB78546	At4g15040	IV, 07545855	+	0	NS		
BAB09764	At5g59190	V, 23599761	+	0/+	M		
AAF25830	At4g00230	IV, 00092935	+	28	S	XSP1	Zhao et al. (2000)
CAB82927	At5g03620	V, 00918737	+	22	S		
AAF79898	At1g20150	I, 06987332	+	22	S		
AAF79897	At1g20160	I, 06990852	+	37	S		
AAF31278	At1g32940	I, 11937750	–	0/+	S		
AAF31277	At1g32950	I, 11941554	+	17	S		
AAF31276	At1g32960	I, 11945467	+	0/+	S		
AAF31279	At1g32970	I, 11948837	–	0	S		
CAB78174	At4g10510	IV, 05460429	+	0	M		
CAB78175	At4g10520	IV, 05464268	+	0	S		
CAB78176	At4g10530	IV, 05473074	+	0	S		
CAB78177	At4g10540	IV, 05476989	+	0	S		
CAB78178	At4g10550	IV, 05481087	+	0	S		
CAB81270	At4g21630	IV, 10456747	+	1500*	S		
CAB81271	At4g21640	IV, 10461333	+	306*	C		
CAB81272	At4g21650	IV, 10465813	+	141	NS		
AAG51763	At1g66210	I, 24317647	+	27	S		
AAG51764	At1g66220	I, 24322448	+	0	S		
CAB87667	At5g11940	V, 03849279	+	0	S		
CAB79488	At4g26330	IV, 12284905	+	0	NS		
BAB09207	At5g45640	V, 18221396	+	0	S		Contains no active site Asp
BAB09208	At5g45650	V, 18227427	+	0/+	S		
AAD12260	At2g04160	II, 01400446	+	108	S	AIR3	Neuteboom et al. (1999)
BAB08348	At5g59810	V, 23810801	+	18	NS		
NP567624	At4g21323	IV, 10304989	+	0			
NP567625	At4g21326	IV, 10311986	+	18			
S8-2							
AAK25995	At5g67360	V, 26586098	+	1346	S	ARA12	Ribeiro et al. (1995)
BAB02339	At3g14067	III, 4658429	+	444	S		
AAC95169	At2g05920	II, 02268826	+	236	S		
BAB11244	At5g51750	V, 20734172	+	892	S		
BAB01030	At3g14240	III, 4741639	+	1124	C		
CAB80215	At4g34980	IV, 15621420	+	1251	S	SLP2	Golldack et al. (2003)
NP563701	At1g04110	I, 01061458	+	18*	S	SDD1	Berger and Altmann (2000)
AAF76468	At1g01900	I, 00310383	+	93	S		
BAB10943	At5g67090	V, 26488017	+	105	S		
S8-3							
BAB09160	At5g44530	V, 17651838	+	25	S		
CAB79043	At4g20430	IV, 09982154	+	88	S		
AAD25747	At1g30600	I, 10841460	+	111	S		
AAD12040	At2g19170	II, 08262703	+	119	S	SLP3	Golldack et al. (2003)
CAB80995	At4g30020	IV, 13642745	+	67	S		
AAF70850	At1g62340	I, 22703034	–	0/+	NS	ALE1	Tanaka et al. (2001)

Table 1 (*continued*)

Protein ID[a]	AGI code[b]	Chrom, position	T-DNA[c]	MPSS[d] ppm	iPSORT[e]	Published name	Notes/references
Not included in phylogenetic tree							
NP197476	At5g19660	V, 06640115	+	126	S		SKI-1.S1P-like

Genes are organized by group numbers S8-1–S8-3, as indicated in Fig. 2A. Genes that occupy the same box under the AGI code are members of genomic clusters. Normalized abundance of transcripts for each gene is reported in parts per million (ppm) and represents the sum of ppm values for all signature sequences, both unique and shared, for each gene. Parts per million values followed by an asterisk are those that include reporting of signature sequences shared by more than one transcript. Parts per million '0' value followed by a '+' indicate the presence of ESTs in GenBank despite the lack of positive MPSS values.

[a] In some cases multiple accession numbers are associated with a given locus. Accession numbers presented are those used for the alignments (Figs. 2–5) and known to possess active site resides, unless otherwise indicated.

[b] AGI ID, Arabidopsis Genome Intitiative code (http://mips.gsf.de/proj/thal/db/index.html).

[c] Availability of T-DNA insert lines was determined from SIGnAL (http://signal.salk.edu/sabout.html).

[d] Normalized transcript abundance obtained from Arabidopsis Massively Parallel Signature Sequencing (MPSS) site (http://dbixs001.dbi.udel.edu/MPSS4/java.html).

[e] Abbreviations for iPSORT results: S, signal sequence; C, chloroplast; M, mitochondria; NS, no signal sequence (i.e., does not contain S, C or M targeting sequences in the N-terminus).

possibly reflecting cell-type-specific roles, the apparent low overall level of S8 family transcripts may be due in part to pathogen- (Tornero et al., 1996) or stress-inducible (Golldack et al., 2003) genes not highly represented among cDNAs used for MPSS. Several S8 (and C1A and A1) family genes not yet detected by MPSS are present among GenBank ESTs and noted in Tables 1–3. Considering the existing evidence that expression of plant subtilisin-like protease genes appears to be regulated by a variety of external and internal cues, *SDD1* and *ALE1* may represent only the beginning in a long list of S8 family members that play important and varied roles in plant growth, development, and defense. Not all of these roles would need to be dependent on peptide bond hydrolysis, however. Apparently, the Asp-His-Ser catalytic triad of serine carboxypeptidase-like (S10 family) sequences can catalyze acyltransferase activities important to glucose polyester biosynthesis in *Lycopersicon pennellii* (Li and Steffens, 2000) and sinapoylmalate biosynthesis in Arabidopsis (Lehfeldt et al., 2000). It has been suggested that the catalytic triad-containing subtilisin-like enzymes should also be considered as potential acyltransferases (Steffens, 2000).

Based on the phylogenetic tree shown in Fig. 2A, we have tentatively divided Arabidopsis subtilisins into three groups, S8-1–S8-3. As was reported recently for F-box proteins from Arabidopsis (Gagne et al., 2002), a comparison of the phylogenetic tree with intron/exon arrangements predicted for the corresponding subtilisin genes yielded very similar subfamily organizations. To simplify the presentation, the gene structures shown (Fig. 2A) are partial and represent only exons including and/or flanked by protease active site residues. The positions within the exons shown for the catalytic Asp, His and Ser residues are conserved for most of the members within each group. The three alternative gene structures are shown mapped to the phylogenetic tree in Fig. 2A using a solid colored line to reflect the Arabidopsis gene structures indicated to the right of the tree. Colored broken lines indicate selected genes from other species with conserved structure. (See Figs. 3–5, for expanded phylogenetic trees including accession numbers, gene names, color-coding and genus names.) The gene structures fit the tree with nearly perfect consistency; the only exception being an outlying gene with the S8-2 structure (BAB10943) which can be considered consistent if the intron-less condition is primitive. The congruence between gene structure and the phylogeny supports the naturalness of our proposed subgroups of known plant subtilisins. The presence in the *Oryza sativa* genome of subtilisins with the S8-1 (BAA89564), S8-2 (AAK63927) and S8-3 (BAB56061) intron/exon arrangements correlated with phylogeny supports the existence of subtilisin genes similar to those in Arabidopsis prior to the divergence of monocots and eudicots.

At 39 genes, the S8-1 group is the largest in Arabidopsis, representing more than half of all Arabidopsis subtilisins described here. There are seven true intron-less genes and two genes containing a single intron each (located at a unique position for each gene) that make up the S8-2 "intron-less" group. The number of S8-2 genes—nine—is consistent with the approximately 20% intron-less genes predicted throughout the Arabidopsis genome (Arabidopsis Genome Initiative, 2000). Like the pathogen-induced tomato subtilisin, *P69A* (Tornero et al., 1996), all tomato genes in the S8-2 group (Fig. 3) are intron-less (Meichtry et al., 1999). To date only a single intron-containing tomato subtilisin, *TMP* (Riggs et al., 2001), has been identified (Fig. 3). This relatively high percentage of intron-less tomato subtilisin genes may be a reflection of incomplete tomato genome sequence data or may be correlated with a species-specific role for intron-less subtilisins. The *SDD1* subtilisin gene (Berger and Altmann, 2000; Von Groll et al., 2002) is an intron-less gene. The smallest group, S8-3, consists of

Table 2
Arabidopsis papain-like cysteine proteases (C1A family)

Protein ID[a]	AGI code	Chrom, position	T-DNA	MPSS ppm	iPSORT	Published name	Notes/references
C1A-1							
BAA02374	At1g47128	1, 16883454	+	3289	S	RD21A	Granulin-like C-term. Koizumi et al. (1993)
BAB08269	At5g43060	V, 16983692	+	45	S		Granulin-like C-term.
CAB80354	At4g36880	IV, 16339175	+	131	S		
BAB02463	At3g19390	III, 06723019	+	508	S		Granulin-like C-term.
BAB02464	At3g19400	III, 06725505	–	214	S		
CAB88124	At3g43960	III, 15783084	–	118	S		
CAB81232	At4g11310	IV, 05848069	–	63*	S		
CAB81233	At4g11320	IV, 05851811	–	1090*	S		
CAB79307	At4g23520	IV, 11238955	+	28	S		
AAK71314	At1g09850	I, 03201852	+	214*	S	XBCP3	Granulin-like C-term. Funk et al. (2002)
C1A-2							
CAB41164	At3g48340	III, 17905734	+	153	S	CEP3	KDEL, Gietl and Schmidt (2001)
CAB41163	At3g48350	III, 17914712	+	13	NS	CEP2	KDEL, Gietl and Schmidt (2001)
BAB09397	At5g50260	V, 20169511	+	72	S	CEP1	KDEL, Gietl and Schmidt (2001)
C1A-3							
AAF25832	At1g20850	I, 07252206	+	468	S	XCP2	Funk et al. (2002)
AAF25831	At4g35350	IV, 15775020	+	139	S	XCP1	Funk et al. (2002)
C1A-4							
AAC49135	At5g45890	V, 18327207	+	0/+	S	SAG12	Gan and Amasino (1995)
C1A-5							
AAF80223	At1g06260	I, 01916450	–	315	S		
AAB67626	At2g34080	II, 14341974	+	151	S		
AAF88126	At1g29080	I, 10157614	+	10	S		
AAF88120	At1g29090	I, 10163223	+	0/+	M		
AAD15594	At2g27420	II, 11674854	+	21*	C		
CAB66413	At3g49340	III, 18302307	+	0	C		
C1A-6							
BAA02373	At4g39090	IV, 17180306	+	3322*	S	RD19A	Koizumi et al. (1993)
AAD23687	At2g21430	II, 09120511	+	3296*	S		
AAK62611	At4g16190	IV, 08136014	–	269	S		
CAB41090	At3g54940	III, 20363364	+	0	S		
C1A-7							
BAB08221	At5g60360	V, 23993950	+	2747	S	AtALEU	Ahmed et al. (2000)
CAB72480	At3g45310	III, 16637664	+	92	S		
C1A-8							
AAK63991	At1g02305	I, 00455778	+	1850*	S		Cathepsin B-like
AAK44008	At4g01620	NF[b]	NF[b]	NF[b]	S		Cathepsin B-like
Not included in phylogenetic tree							
AAF88125	At1g29110	1, 10171669	–	0	S		No active site Cys
AAC24376	At1g02300	1, 00453288	+	0	S		Cathepsin B-like, lacks Gln19 residue that is part of the Cys active site consensus sequence

Genes are organized by group numbers C1A-1–C1A-8, as indicated in Fig. 2B. Genomic clusters and normalized abundance of transcripts are indicated as in Table 1.

[a] Column headings are as described for Table 1.

[b] NF, not found.

six genes and includes the *ALE1* subtilisin (Tanaka et al., 2001).

Numerous examples of genomic clustering of Arabidopsis subtilisins on chromosomes I, III, IV and V are evident from chromosome positions and these are indicated as boxed clusters of genes in Table 1. Together, these clustered genes represent 64% of the subtilisin-like proteases in Arabidopsis. Subtilisin gene clusters exist only within the S8-1 group (Table 1). Subtilisin gene duplication leading to cluster formation may be an important component of functional diversification. New functions for duplicated subtilisins could result from sequence divergence affecting gene regulatory elements or protein targeting. For example, the genomic cluster

Table 3
Arabidopsis pepsin-like aspartic proteases (A1 family)

Protein ID[a]	AGI code	Chrom, position	T-DNA	MPSS ppm	iPSORT	Published name	Notes[b]/references
A1-1							
BAB01116	At3g18490	III, 06349087	+	2654	S		
AAG50814	At1g25510	I, 08959367	–	122	S		
CAB71112	At3g61820	III, 22889033	+	525	S		AAK64003 Contains site I. DTG only
AAF97328	At1g01300	I, 00117065	+	2045	S		
BAB03167	At3g20015	III, 06978739	+	54	C		
CAB96831	At5g10760	V, 03400669	+	26	S		CND41-like
CAB96832	At5g10770	V, 03403329	+	239	M		CND41-like
AAG52249	At1g79720	I, 29650398	+	231	NS		
BAB11161	At5g07030	V, 02183599	+	2063	S		Site I. DTS
CAB81805	At3g54400	III, 20149247	+	163	S		Site I. DTS
AAB60729	At1g09750	I, 03157545	–	844*	S		Site I. DTS
CAB82992	At5g02190	V, 00435320	+	360	S		
AAB87120	At2g39710	II, 00467267	+	82	S		
AAG51309	At1g66180	I, 24299133	+	789	S		
AAL06853	At5g37540	V, 14626772	+	7	S		
NP198319	At5g33340	V, 12370818	+	0	C		
NP198320	At5g33350	V, 12373454	+	0	NS		
AAD38257	At1g64830	I, 23743183	–	11	S		
AAG51267	At1g31450	I, 11259990	+	0	S		
AAD21501	At2g28010	II, 11879121	+	13*	S		
AAC98463	At2g28030	II, 11882750	–	13*	C		
AAC98462	At2g28040	II, 11884745	+	13*	C		
AAD29831	At2g28220	II, 11982495	+	13*	S		Contains 4 DT/SG sites
AAC34482	At2g03200	II, 00965502	+	25	S		
AAL06828	At4g16560	IV, 08289255	–	0/+	S		
BAB09497	At5g45120	V, 17954910	+	0	S		
AAL14384	At3g52500	III, 19474600	+	1666*	S		
AAD21712	At2g42980	II, 17823548	–	12	NS		
AAL11556	At3g59080	III, 21845773	+	108	S		
BAB03090	At3g25700	III, 09360165	+	78	S		
BAB02414	At3g12700	III, 04037211	+	239	NS		
CAB43838	At4g30030	IV, 13646704	+	0	S		
CAB43839	At4g30040	IV, 13650096	+	17	S		Site I. DTA
NP565559	At2g23945	II, 10192309		0			
CAB45491	At4g12920	IV, 06532767	–	23	M		
A1-2							
BAB10606	At5g22850	V, 07609266	+	146	S		
AAF18253	At1g08210	I, 02577118	+	199	C		
AAD20143	At2g36670	II, 15313695	+	160	S		
CAB85978	At3g42550	III, 14674958	–	0	S		
AAD26876	At1g65240	I, 23882875	+	13	S		
BAB09366	At5g36260	V, 13998978	+	15	S		
AAF26986	At3g02740	III, 00590570	+	53	S		
AAF29389	At1g05840	I, 01762844	+	140	S		
CAB62655	At3g51330	III, 19062434	+	0/+	S		AAK74041 Contains site II. DTG only
CAB62656	At3g51340	III, 19065967	+	31	NS		
AAB80784	At2g17760	II, 07662232	+	106	S		
CAA21474	At4g35880	IV, 15957828	+	59	S		
CAB92049	At5g10080	V, 03150841	+	28	S		
AAG51657	At1g77480	I, 28768087	+	123	M		Nucellin-like
AAG50556	At1g44130	I, 16317182	+	92	S		Nucellin-like
CAB38807	At4g33490	IV, 15073270	–	39	S		Nucellin-like
AAK25892	At1g49050	I, 17738200	+	161	NS		Nucellin-like
A1-3							
BAB08274	At5g43100	V, 17013172	+	162	S		
CAB62113	At3g50050	III, 18563096	+	80	S		

(*continued on next page*)

Table 3 (*continued*)

Protein ID[a]	AGI code	Chrom, position	T-DNA	MPSS ppm	iPSORT	Published name	Notes[b]/references
A1-4							
AAB60773	At1g62290	I, 22662018	+	40	S	Pasp-A2	Chen et al. (2002) Gruis et al. (2002)
AAK50111	At4g04460	IV, 02224236	+	1360	S	Pasp-A3	Chen et al. (2002)
AAL08259	At1g11910	I, 04017119	+	876	S	Pasp-A1	Mutlu et al. (1998) Chen et al. (2002)
A1-5							
CAA18108	At4g22050	IV, 10648364	+	0/+	S		
AAF27055	At1g69100	I, 25631903	+	0	S		
Not included in phylogenetic tree							
CAB62657	At3g51350	III, 19069439	+	0	S		Contains site I DTG only
CAB62658	At3g51360	III, 19073248	+	0	S		Contains site II DTG only

Genes are organized by group numbers A1-1–A1-5, as indicated in Fig. 2C. Genomic clusters and normalized abundance of transcripts are indicated as in Table 1.

[a] Column headings are as described for Table 1.

[b] Deviations from the DT/SG active site motif are noted.

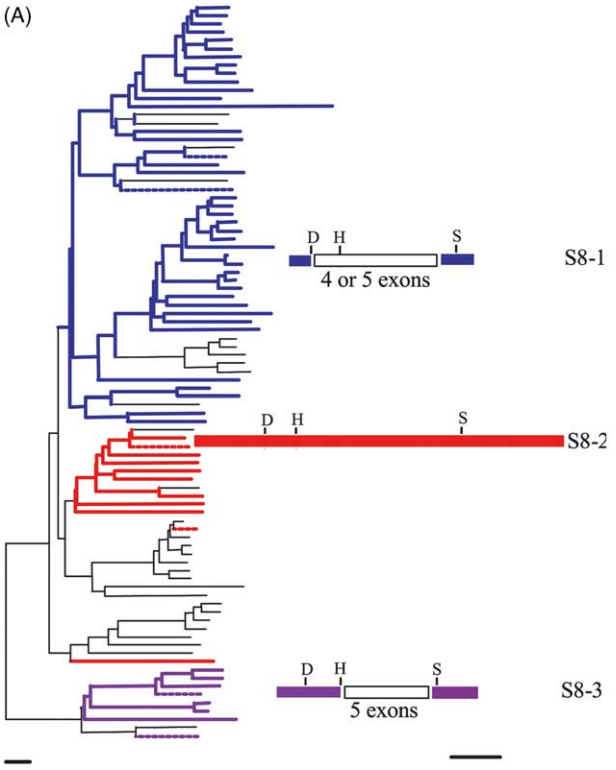
of the three expressed S8-1 genes CAB81270, CAB81271 and CAB81272 encodes proteins predicted to localize to the secretory pathway, the chloroplast and the cytosol. Functional specialization could also arise from changes in protease-substrate interaction mediated by the PA domain (Mahon and Bateman, 2000) or evolution of the catalytic triad as an acyltransferase (Steffens, 2000).

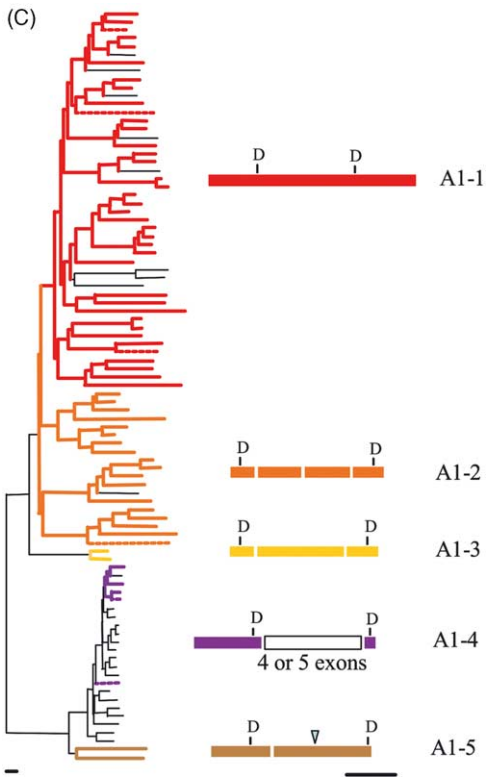
2.2. C1A proteases

The C1 family of proteases includes the papain subfamily (C1A) and the bleomycin hydrolase subfamily (C1B). Although sequences predicting C1 family proteases are found in all kingdoms, plant C1 sequences appear to be limited to representatives from the A subfamily (MEROPS). The multigene family of proteases presented in Table 2 includes 30 proteases homologous to the C1A family of papain-like cysteine proteases. All Arabidopsis C1A enzymes except AAK44008 and AAK63991 (Table 2) possess an N-terminal prodomain containing the non-contiguous Glu/Asp-Arg-Phe/Tyr/Leu-Asn-Ile/Ala/Val-Asn/Gln (ERFNIN) signature found in human cathepsin L. This feature has been used to distinguish the cathepsin L-like proteases from the cathepsin B-like proteases, which lack the ERFNIN signature (Karrer et al., 1993). Hence AAK44008 and AAK63991 have been designated as cathepsin B-like proteases. Four of the Arabidopsis C1A proteases as well as several papain-like enzymes from other plant species (Gietl et al., 2000) contain cysteine-rich, granulin-like (Bateman and Bennett, 1998) C-terminal extensions (Table 2). Granulin is an 8.8-kD polypeptide capable of mediating the growth of human cells in certain experimental models (e.g., Lu and Serrero, 2001). Plant granulin-like domains of C1A proteases may participate in the regulation of protease solubility and activation (Yamada et al., 2001). Table 2 includes two potential C1A family pseudogenes, AFF88125 and AAC24376, both of which lack amino acid residues associated with the cysteine active site and are not represented among reported ESTs.

The papain-like enzymes from plants have been described as catalysts of protein remobilization during seed germination and organ senescence and they have been implicated in numerous plant cell suicide programs (Gan and Amasino, 1995; Beers et al., 2000; McCann et al., 2000; Gietl and Schmid, 2001; Hayashi et al., 2001; Funk et al., 2002; Wan et al., 2002). Papain-like enzymes may also be important in defense against pathogens (Solomon et al., 1999) and insects (Visal et al., 1998; Pechan et al., 2000) and as catalysts of wound sealing (El Moussaoui et al., 2001). Recently, a C1A protease from tomato, Rcr3 (AAM19208), was shown to be required for the function of the disease resistance gene *Cf-2* (Kruger et al., 2002). The precise nature of the requirement for *Rcr3* is not yet known.

Of the three subordinate classes of proteases reported here, the Arabidopsis C1A family includes the highest number of partially characterized genes (Table 2). SAG12 is a senescence-specific protease (Gan and Amasino, 1995). Leaf senescence is also associated with increased levels of *SAG2*/*AtALEU* mRNA (Hensel et al., 1993). Vacuolar targeting of *AtALEU* is mediated by the vacuolar sorting receptor AtELP through the N-terminal propeptide Asn-Pro-Ile-Arg (NPIR) domain of *AtALEU* (Ahmed et al., 2000). Drought stress and senescence are associated with increased expression of *RD21A* (Koizumi et al., 1993; Yamada et al., 2001). Salt stress induces the fusion of ER bodies containing RD21A with vacuoles where RD21A may be processed to become proteolytically active (Hayashi et al., 2001). The cDNA for XBCP3 was cloned from a bark cDNA library (Zhao et al., 2000), and in leaves and stems the





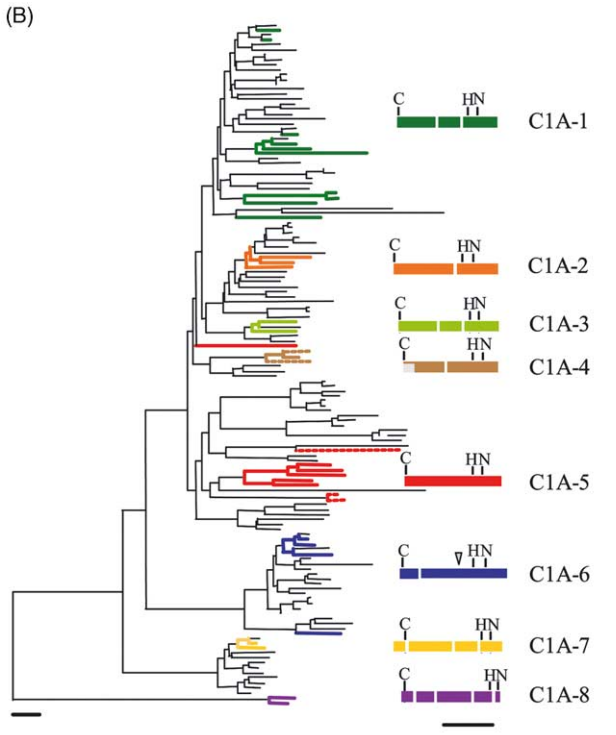


Fig. 2. Phylogenetic tree of (A) 89 plant S8 proteases (B) 138 plant CIA proteases and (C) 92 plant A1 proteases and gene structure diagrams for Arabidopsis S8 CIA and A1 proteases. (A) This is a neighbor-joining tree based on PAM distances generated from an alignment of protein sequences (based on the program ClustalW, and prodist and neighbor of PHYLIP). Tentative group assignments, S81-S83, are indicated by color-coding using solid lines for Arabidopsis sequences and broken lines for selected members from other species. The intron/exon arrangement (available from http://mips.gsf.de/proj/thal/db/index.html) for the majority of members in each group is shown to the right of the tree in corresponding color. Translated exons are drawn to scale except where grouped together (represented by a single, unfilled box, as for S8-1 and S8-3 sequences). To simplify the diagram, only exons flanked by and containing catalytic residues are shown. Positions for the catalytic amino acid residues Asp (D), His (H) and Ser (S) are indicated for each group. The bar on the left represents a tree branch length equivalent to 0.1 amino acid changes per residue. The bar on the right is equal to 100 amino acids in the gene structure diagrams. Expanded versions of the tree with accession numbers, gene names and genus names for all members can be found in Fig. 3. (B) Phylogenetic tree analysis was performed as for (A). Eight tentative group assignments (C1A-1–C1A-8) are indicated by color-coding using solid lines for Arabidopsis sequences and broken lines for selected members of the tree from other species. The intron/exon diagrams were prepared as for (A). Positions for the catalytic amino acid residues Cis (C), His (H) and Asn (N) are indicated for each group. The blue triangle above the group C1A-6 gene points to the position of an additional intron for one member. The scale bars are as indicated for (A). Expanded versions of the tree can be found in Fig. 4. (C) Phylogenetic tree analysis was performed as for (A). Five tentative group assignments (A1-1–A1-5) are indicated by color-coding using solid lines for Arabidopsis sequences and broken lines for selected members of the tree from other species. The intron/exon diagrams were prepared as for (A). Exons are drawn to scale except where grouped together (represented by a single, unfilled box, as for A1-4 sequences). Positions for the two catalytic Asp (D) residues are indicated. The blue triangle above the group A1-5 diagram points to the position of an additional intron for one member. The scale bars are as indicated for (A). Expanded versions of the tree can be found in Fig. 5.

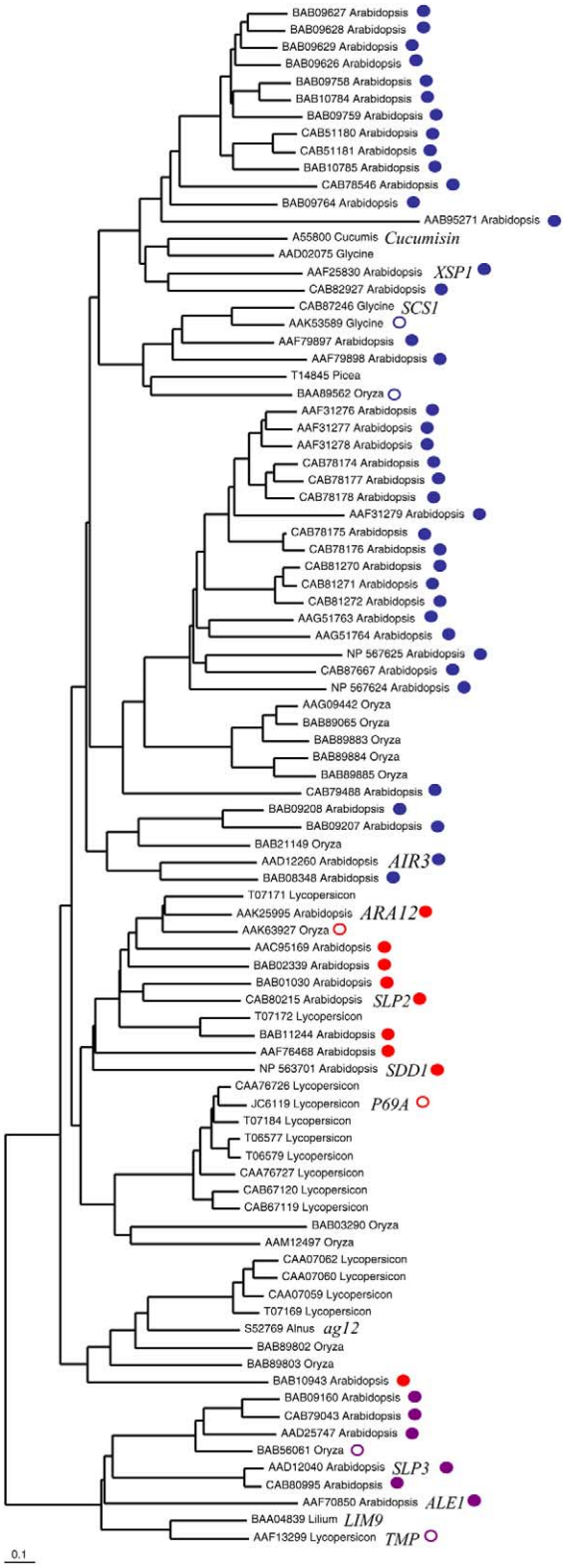


Fig. 3. Expanded phylogenetic tree of 89 plant S8 proteases showing accession numbers, gene names, group color-coding and genus names. Phylogenetic tree analysis was performed as for Fig. 2A. Color-coding is the same as for Fig. 2A. Colored, filled circles are used at branch tips of Arabidopsis sequences and open circles are used at branch tips where similar gene structure was noted for representative sequences from other species. The bar represents the branch length equivalent to 0.1 amino acid changes per residue.

expression of an *XBCP3* promoter-*GUS* reporter gene is limited to hydathodes and vascular tissues (Funk et al., 2002). Preliminary evidence links the three KDEL-containing C1A proteases, CEP1, CEP2 and CEP3, with programmed cell death (Gietl and Schmid, 2001). Promoters for *XCP1* and *XCP2* direct GUS expression in tracheary elements, where XCP1 localizes to vacuoles (Funk et al., 2002). CAB41090 and CAB81232 are reported to be seed-specific (Gruis et al., 2002). Transcripts for 25 of the 30 C1A proteases are present among the cDNA libraries used for MPSS (Table 2). Of these expressed genes, six are highly expressed and three of these highly expressed genes, *RD21A*, *RD19A* and *AtALEU*, have been partially characterized (Koizumi et al., 1993; Ahmed et al., 2000; Hayashi et al., 2001), although their precise functions are still unknown.

The Arabidopsis C1A protease family includes the most highly structured phylogenetic tree with eight different groups designated C1A-1 to C1A-8 (Fig. 2B). Distinct intron/exon arrangements are noted for each major branch of the tree, except for the C1A-1 and C1A-3 groups whose members exhibit the same gene structure. Hence, as noted for the S8 proteases, C1A proteases can be organized in similar ways using either intron/exon arrangements or amino acid phylogeny. The two C1A-3 proteases (AAF25832 and AAF25831) are the only xylem-specific C1A proteases in Arabidopsis (C. Zhao, E. Beers and A. Dickerman, unpublished). It would be interesting to determine whether orthologous genes are present in non-vascular plants. Although there are no intron-less genes among the Arabidopsis C1A genes, six genes (C1A-5 group) include all three active site residues within a single exon. Genes for the tomato C1A proteases AAM19208 and AAM19207 (See Fig. 4) also exhibit the C1A-5 structure. AAM19207 is encoded by *Rcr3*, a gene required for Cf-2-dependent disease resistance (Kruger et al., 2002). SAG12, the only Arabidopsis protease shown to be restricted to leaf senescence, is the sole Arabidopsis member of the C1A-4 group. The C1A-2 group includes all three KDEL-containing papain-like enzymes. Two aleurain-like proteases are present (C1A-7). The two most divergent papain-like enzymes are the two cathepsin B-like proteases in the C1A-8 group. Clustering of genes for papain-like enzymes is limited to two-gene clusters on chromosomes I, III and IV, and these are grouped together in boxes in Table 2.

2.3. A1 proteases

A1 family protease sequences are known only from eukaryotes (MEROPS). Fifty-nine A1 proteases were identified among the annotated Arabidopsis genes for this report. Aspartic proteases from other families predicted from the Arabidopsis genome include the large (approximately 45-member) A11 family of

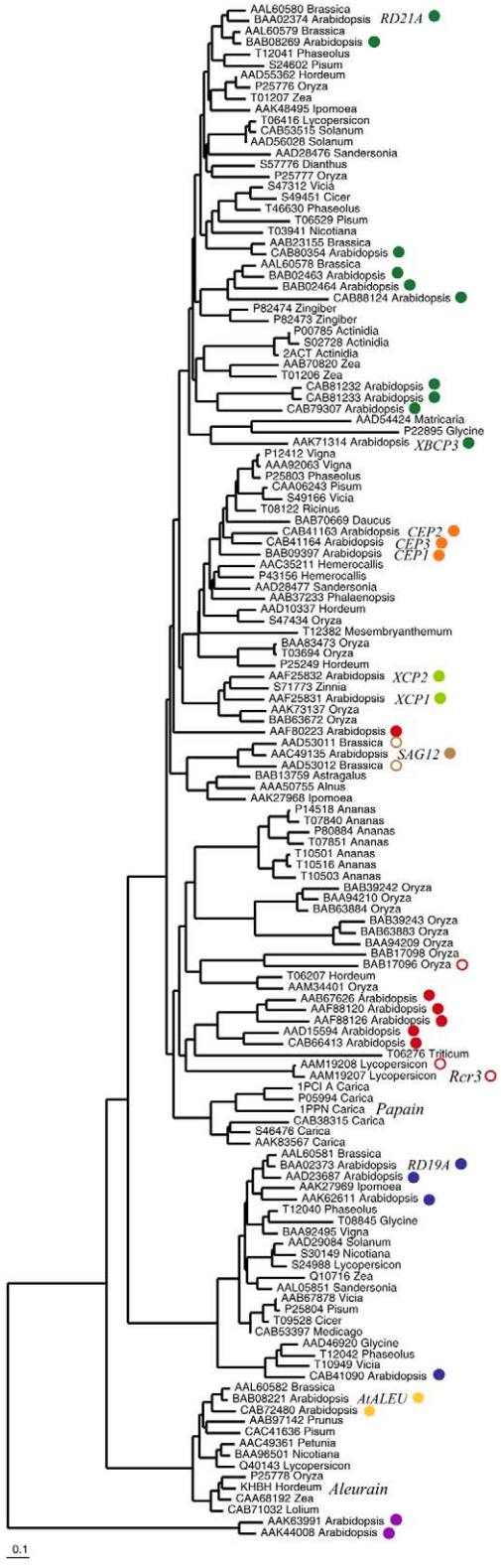


Fig. 4. Expanded phylogenetic tree of 138 plant C1A proteases showing accession numbers, gene names, group color-coding and genus names. Phylogenetic tree analysis was performed as for Fig. 2A. Color-coding is the same as for Fig. 2B. Colored, filled circles are used at branch tips of Arabidopsis sequences and open circles are used at branch tips where similar gene structure was noted for representative sequences from other species. The bar represents the branch length equivalent to 0.1 amino acid changes per residue.

retrotransposon endopeptidases and two presenilin-like proteins of family A22 (AAL24266 and AAD23630). Human presenilins are central to Alzheimer's disease as the putative catalytic component of the γ-secretase required for the release of the amyloid-β peptide from the amyloid-β precursor protein (Esler and Wolfe, 2001). Roles for plant presenilin-like proteins have not yet been reported.

Compared to the S8 and C1A plant proteases, fewer A1 proteases have been studied. A1 proteases may be important to disease resistance and programmed cell death. Activation tagging of an as yet unidentified Arabidopsis A1 protease led to enhanced resistance to virulent strains of *Pseudomonas* and was correlated with constitutive expression of defense-related genes and elevated salicylic acid levels (Y. Xia, C. Lamb and R. Dixon, personal communication). Four A1 proteases have been classified as nucellin-like (Table 3) due to their homology to barley nucellin (T06213), an A1 protease associated with post-anthesis degeneration of nucellar tissue (Chen and Foolad, 1997). Several Arabidopsis A1 peptidases are annotated as "chloroplast nucleoid DNA binding protein-like" (CND41-like). Originally identified from tobacco, CND41 (T01996) exhibits both proteolytic (Murakami et al., 2000) and chloroplast DNA-binding (Nakano et al., 1997) capabilities in vitro. DNA binding was dependent on a Lys-rich region located N-terminal to the site I DT/SG protease active site motif. Like phytepsin and its mammalian counterpart pepsin, CND41 is most active at acidic pH (pH 2–4). In contrast to pepsin, proteolytic activity of CND41 is only slightly sensitive to the pepsin inhibitor pepstatin (Murakami et al., 2000). CND41 is inhibited by NADPH, Fe^{3+}, and nucleotide triphosphates in vitro, potentially reflecting mechanisms for coordinating CND41-mediated proteolysis with chloroplast metabolism.

Pasp-A1 and two other Arabidopsis A1 proteases, *Pasp-A2* and *Pasp-A3* (Table 3), have been classified as phytepsins based on their exceptional similarity to mammalian enzymes pepsin and cathepsin D. *Pasp-A1,* has been implicated in lectin processing in seeds (Mutlu et al., 1998). Yeast counterparts to phytepsin are also known, e.g. the *Saccharomyces cerevisiae PEP4* gene product, proteinase A. In *S. cerevisiae*, proteinase A is required for the activation of several vacuolar zymogens (van den Hazel et al., 1992). Whether aspartic proteases are required for activation of plant vacuolar zymogens is not known. Phytepsins are distinguished from the mammalian and yeast enzymes by a saposin-like, plant-specific insert (approximately 100 amino acids) C-terminal to the second DT/SG (Kervinen et al., 1999) (Fig. 1C). For barley phytepsin, the saposin-like domain is required for vacuolar localization and also appears to influence the ER export process (Tormakangas et al., 2001). Enzymatic activity is not dependent on the saposin-like domain, however. The three

Arabidopsis phytepsins are differentially expressed: *Pasp-A2*, primarily in seeds (Chen et al., 2002; Gruis et al., 2002), *Pasp-A1*, in all organs, and *Pasp-A3*, primarily in flowers (Chen et al., 2002). Several putative A1 family genes exhibit deviations from the bilobed DT/SG consensus (Table 3). Where possible, i.e., for BAB11161, CAB81805 and AAB60729, these deviations were confirmed by conceptual translation of available ESTs. A pair of partial A1 genes, CAB62657 and CAB62658, each containing only one of the active site lobes, is included in Table 3. Neither gene is represented among the transcripts in the MPSS or other EST data sets, suggesting that these are pseudogenes. Interestingly, two A1 peptidases (CAB62655 and CAB71112) predicted to possess two DT/SG sites are in fact transcribed as mRNA (AAK74041 and AAK64003, respectively) encoding only a single DT/SG site each (Table 3). Perhaps some single DT/SG-site proteins function as homo or heterodimers to form a bilobed functional aspartic protease. A search for A1 protease transcripts revealed that 48 (81%) of the predicted A1 peptidases are represented in the MPSS data set. Of the five highly expressed genes, only one, *Pasp-A1*, has been the subject of biochemical and molecular studies (D'Hondt et al., 1997; Chen et al., 2002).

Protein sequence divergence is higher in the A1 family than in the S8 or C1A families, as shown by the distance scale bar in the phylogenetic tree of Fig. 2C. Tentative group assignments (A1-1–A1-5) have been made for Arabidopsis A1 proteases based on the phylogeny. Fig. 2C is color coded to show members of each group and a representative gene structure is diagrammed to show the most common intron/exon arrangement (flanked by the active site Asp residues) within each group. As noted for the S8 and C1A families, the A1 proteases compiled here can be grouped in similar ways using either gene structure or amino acid phylogeny. The expanded phylogenetic tree including gene names, color-coding, accession numbers and genus names for all sequences is shown in Fig. 5. Group A1-1 is the largest group at 35 sequences of which 76% are true intron-less genes. Of the remaining A1-1 genes, one (AAL06828) is predicted to include 21 introns and the remainder, distributed throughout the A1-1 group, possess one or two introns each. Structural similarity among the A1-1 genes with predicted introns was not apparent. The second largest group, A1-2, includes the nucellins (Table 3). The A1-4 sequences are most closely related to barley phytepsin (P42210) and are the only Arabidopsis A1 proteases that contain the saposin-like domain. As reported here for several groups within the C1A and S8 families, some A1 family groups (A1-1, A1-2 and A1-4) exhibit conservation of intron/exon organization between monocots (*Oryza* and *Hordeum*) and Arabidopsis. The two Arabidopsis A1 proteases most similar to CND41, CAB96831 and CAB96832, are

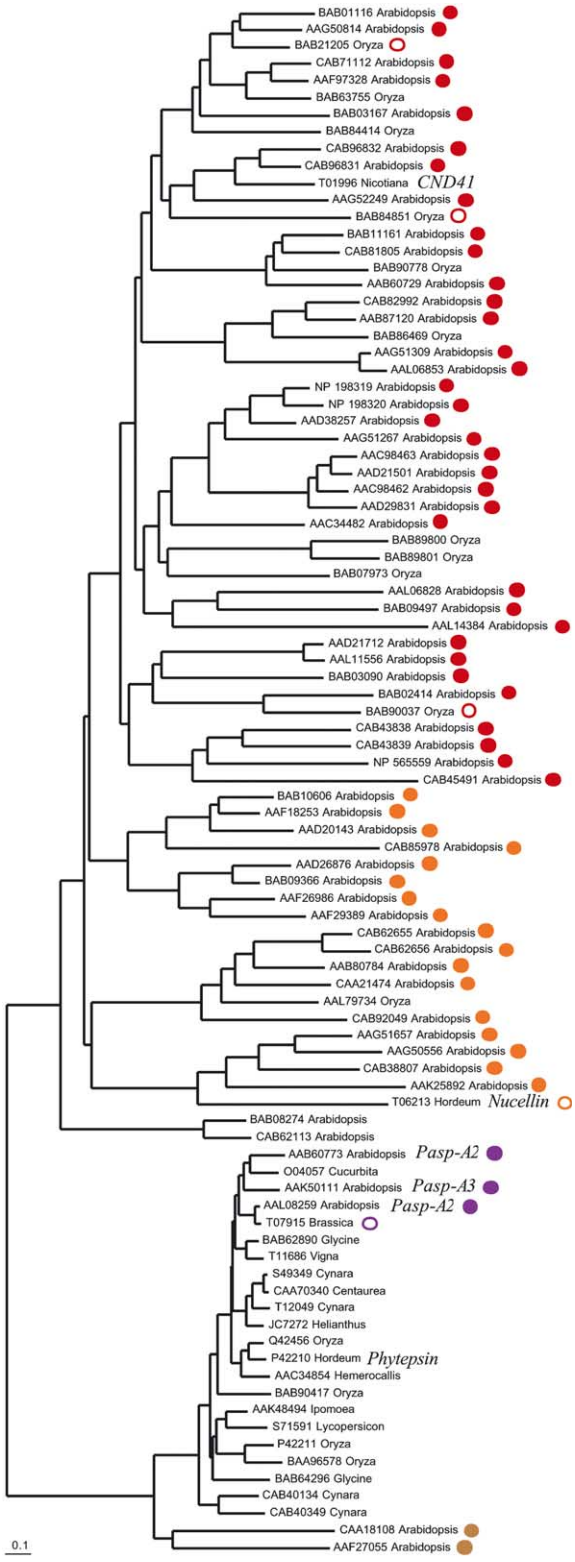


Fig. 5. Expanded phylogenetic tree of 91 A1 proteases showing accession numbers, gene names, group color-coding and genus names. Phylogenetic tree analysis was performed as for Fig. 2A. Color-coding is the same as for Fig. 2C. Colored, filled circles are used at branch tips of Arabidopsis sequences and open circles are used at branch tips where similar gene structure was noted for representative sequences from other species. The bar represents the branch length equivalent to 0.1 amino acid changes per residue.

clustered on chromosome V. This and additional gene clusters of A1 proteases are presented in boxes in Table 3.

3. Concluding remarks

During the past decade it has become apparent that targeted proteolysis mediated by the ubiquitin/ 26S proteasome pathway is important to many aspects of plant biology (Vierstra, 2003). These findings combined with recently revealed novel roles attributed to two subtilisin-like genes, *SDD1* (Berger and Altmann, 2000; Von Groll et al., 2002), *ALE1* (Tanaka et al., 2001) and a papain-like gene *Rcr3* (Kruger et al., 2002), have broadened our perception of plant proteolytic systems to include roles far beyond developmentally programmed and inducible nitrogen recycling and autolysis associated with organ senescence and programmed cell death. Moreover, recent revelations that some predicted serine proteases possessing catalytic triads function not as hydrolases but as acyltransferases of secondary metabolism (Li and Steffens, 2000; Steffens, 2000) suggest that some of the subtilisin-like sequences considered here may code for enzymes not involved in peptide bond hydrolysis.

We report several examples of structural conservation between S8, C1A and A1 genes from rice, barley, tomato and soybean compared with those from Arabidopsis, indicating that some common, essential plant protease roles were established before the divergence of monocots and eudicots. In fact, phylogenetic trees that included non-plant protease sequences (data not shown) isolated multiple plant groups in both the C1A and A1 protease families, separated by intervening non-plant clades. Though the robustness of these isolated plant lineages was not tested, these groups remained as the major distinct branches in later plant-only analyses (Figs. 2–5), suggesting that the depth of divergences within the cysteine and aspartate protease groups dates at least to before the divergence of metazoans. Structural similarity reported here for protease genes across monocot and eudicot species suggests that a comprehensive functional analysis of these protease families in Arabidopsis is a realistic strategy for understanding protease function in a variety of important crop species.

The existence of numerous protease gene clusters within the S8, C1A and A1 families hints at tandem gene duplications preceding sequence divergence leading to functional specialization. Recent demonstrations that serine proteases possessing catalytic triads can catalyze acyltransferase reactions hint at possible non-proteolytic functions (at least for the S8 family enzymes). In addition to potential recruitment of some genes for non-hydrolytic activities, restricted gene expression may be an important element of functional specialization of closely related proteases (Chapman et al., 1997). The evidence summarized here illustrating the variety of cell type-, tissue-specific, developmentally-regulated and inducible expression patterns exhibited by the plant proteases suggests that we are only beginning to appreciate the degree of specialization occurring within the multigene families of C1A, S8, and A1 proteases.

4. Experimental

4.1. Alignment, HMM building, and phylogenetic tree inference

For each protease family, a set of known Arabidopsis examples was assembled manually and aligned using ClustalW version 1.8 (Thompson et al., 1994). This initial alignment was trimmed on the left and right sides to remove regions where alignment was poor due to the variable nature of the processed leader and downstream regions. This trimming was based on sequence conservation at each position in a sliding window moving from each edge towards the center, trimming at the position where the score exceeded a specified, but empirically adjusted, value. The trimmed sequences were subjected to another round of alignment. This final alignment was used to generate a profile hidden Markov model (HMM) using the program hmmbuild from the HMMER 2.2 suite developed by the Sean Eddy's group (http://hmmer.wustl.edu/). HMMs were built specifying the "local alignment" model in order to identify partial matches to the pattern caused by truncated sequences or protein domain re-arrangement. Each of the three resulting HMMs, one for each protease family, was used to search (using HMMER's hmmsearch) the non-redundant protein set known as "nr" available from Genbank which combines Genpept, Swissprot, and Tremble and is essentially comprehensive of all available protein sequences, putative or otherwise. This resulted in a large set of related proteins from other plant species as well as non-plants for cysteine and aspartic proteases. These large sequence sets were scanned for additional Arabidopsis sequences not encountered in the initial set, with only one new member found. Cursory phylogenetic analysis using the protdist and neighbor programs of PHYLIP 3.6 developed by Joe Felsenstein (http://evolution.genetics.washington.edu/phylip.html) was used to visualize the broad relationship of plant clades to those of other groups (data not shown). Non-redundant sets of plant homologs for each protease family were assembled from the results of the HMM search and aligned and trimmed as described above. The aligned protein sequences were used to generate PAM distance matrices using the protdist program of PHYLIP, which was then used to generate a neighbor-joining tree using the neighbor program of PHYLIP.

4.2. Gene structure analysis

Analyses of gene structure (intron/exon arrangements) were made by hand using unspliced sequences for all genes obtained from the MIPS Arabidopsis thaliana Genome Database (MATDB) (Schoof et al., 2002) (http://mips.gsf.de/proj/thal/db/index.html). To simplify the analysis, only the exons and introns flanked by or containing active site residues were considered for inclusion in the schematic representation of intron/exon arrangements presented in Fig. 2A–C.

4.3. Transcript abundance and T-DNA identification

Normalized transcript abundance in parts per million (ppm) was determined by adding ppm values from all signature sequences (both unique and shared) for a given gene as reported by the Arabidopsis Massively Parallel Signature Sequencing (MPSS) site (http://dbixs001.dbi.udel.edu/MPSS4/java.html). Protease genes tagged by T-DNA were identified from among those listed by SIGnAL (as of 5/21/03).

Acknowledgements

This work was supported by the US Department of Agriculture-National Research Initiative Competitive Grants Program (project No. 9801401 to E.P.B.), the US Department of Agriculture-Cooperative State Research, Education and Extension Service (project No. 2001-34448-10462 to A.W.D.) and by the National Science Foundation (grants No. IBN-0131386 to E.P.B. and No. IBN-9807801 to A.M.J). E.P.B. thanks Dr. Susannah Gal (SUNY, Binghamton) for critical reading of the manuscript and Ms. Maura Wood (Virginia Tech) for technical assistance. A.M.J. thanks Mr. Brian Jones (UNC) and Dr. Michael Purugganan (NCSU) for technical assistance.

References

Ahmed, S.U., Rojo, E., Kovaleva, V., Venkataraman, S., Dombrowski, J.E., Matsuoka, K., Raikhel, N.V., 2000. The plant vacuolar sorting receptor AtELP is involved in transport of $NH(_2)$-terminal propeptide-containing vacuolar proteins in *Arabidopsis thaliana*. Journal of Cell Biology 149, 1335–1344.

Arabidopsis Genome Initiative, 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408, 796–815.

Batchelor, A.K., Boutilier, K., Miller, S.S., Labbe, H., Bowman, L., Hu, M., Johnson, D.A., Gijzen, M., Miki, B.L., 2000. The seed coat-specific expression of a subtilisin-like gene, SCS1, from soybean. Planta 211, 484–492.

Bateman, A., Bennett, H.P., 1998. Granulins: the structure and function of an emerging family of growth factors. Journal of Endocrinology 158, 145–151.

Beers, E.P., Zhao, C., 2001. Arabidopsis as a model for investigating gene activity and function in vascular tissues. In: Morohoshi, N., Komamine, A. (Eds.), Molecular Breeding of Woody Plants. Proceedings of the International Wood Biotechnology Symposium, Narita, Japan. Elsevier, NY, pp. 43–52.

Beers, E.P., Woffenden, B.J., Zhao, C., 2000. Plant proteolytic enzymes: possible roles during programmed cell death. Plant Molecular Biology 44, 399–415.

Berger, D., Altmann, T., 2000. A subtilisin-like serine protease involved in the regulation of stomatal density and distribution in *Arabidopsis thaliana*. Genes and Development 14, 1119–1131.

Brenner, S., Johnson, M., Bridgham, J., Golda, G., Lloyd, D.H., Johnson, D., Luo, S., McCurdy, S., Foy, M., Ewan, M., Roth, R., George, D., Eletr, S., Albrecht, G., Vermaas, E., Williams, S.R., Moon, K., Burcham, T., Pallas, M., DuBridge, R.B., Kirchner, J., Fearon, K., Mao, J., Corcoran, K., 2000. Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead assays. Nature Biotechnology 18, 630–634.

Chapman, H.A., Riese, R.J., Shi, G.P., 1997. Emerging roles for cysteine proteases in human biology. Annual Review of Physiology 59, 63–88.

Chen, F., Foolad, M.R., 1997. Molecular organization of a gene in barley which encodes a protein similar to aspartic protease and its specific expression in nucellar cells during degeneration. Plant Molecular Biology 35, 821–831.

Chen, X., Pfeil, J.E., Gal, S., 2002. The three typical aspartic proteinase genes of *Arabidopsis thaliana* are differentially expressed. European Journal of Biochemistry 269, 4675–4684.

D'Hondt, K., Stack, S., Gutteridge, S., Vandekerckhove, J., Krebbers, E., Gal, S., 1997. Aspartic proteinase genes in the Brassicaceae *Arabidopsis thaliana* and *Brassica napus*. Plant Molecular Biology 33, 187–192.

El Moussaoui, A., Nijs, M., Paul, C., Wintjens, R., Vincentelli, J., Azarkan, M., Looze, Y., 2001. Revisiting the enzymes stored in the laticifers of *Carica papaya* in the context of their possible participation in the plant defense mechanism. Cellular and Molecular Life Sciences 58, 556–570.

Elagoz, A., Benjannet, S., Mammarbassi, A., Wickham, L., Seidah, N. G., 2001. Biosynthesis and cellular trafficking of the convertase SKI-1/S1P: ectodomain shedding requires SKI-1 activity. Journal of Biological Chemistry 277, 11265–11275.

Esler, W.P., Wolfe, M.S., 2001. A portrait of alzheimer secretases—new features and familiar faces. Science 293, 1449–1454.

Estelle, M., 2001. Proteases and cellular regulation in plants. Current Opinion in Plant Biology 4, 254–260.

Fu, X., Richards, D.E., Ait-Ali, T., Hynes, L.W., Ougham, H., Peng, J., Harberd, N.P., 2002. Gibberellin-mediated proteasome-dependent degradation of the barley DELLA protein SLN1 repressor. The Plant Cell 14, 3191–3200.

Funk, V., Kositsup, B., Zhao, C., Beers, E.P., 2002. The Arabidopsis xylem peptidase XCP1 is a tracheary element vacuolar protein that may be a papain ortholog. Plant Physiology 128, 84–94.

Gagne, J.M., Downes, B.P., Shui, S.H., Durski, A.M., Vierstra, R.D., 2002. The F-box subunit of the SCF E3 complex is encoded by a diverse superfamily of genes of Arabidopsis. Proceedings of the National Academy of Sciences of the United States of America 99, 11519–11524.

Gan, S., Amasino, R.M., 1995. Inhibition of leaf senescence by autoregulated production of cytokinin. Science 270, 1986–1988.

Gietl, C., Schmid, M., 2001. Ricinosomes: an organelle for developmentally regulated programmed cell death in senescing plant tissues. Naturwissenschaften 88, 49–58.

Gietl, C., Schmid, M., Simpson, D., 2000. Ricinosomes and aleurain-containing vacuoles (ACVs): protease-storing organelles. In: Robinson, D.G., Rogers, J.C. (Eds.), Vacuolar Compartments. Annual Plant Review, Vol. 5. Sheffield Academic Press, Sheffield, UK, pp. 90–111.

Glathe, S., Kervinen, J., Nimtz, M., Li, G.H., Tobin, G.J., Copeland, T.D., Ashford, D.A., Wlodawer, A., Costa, J., 1998. Transport and activation of the vacuolar aspartic proteinase phytepsin in barley (*Hordeum vulgare* L.). Journal of Biological Chemistry 273, 31230–31236.

Golldack, D., Vera, P., Dietz, K.J., 2003. Expression of subtilisin-like serine proteases in *Arabidopsis thaliana* is cell-specific and responds to jasmonic acid and heavy metals with developmental differences. Physiologia Plantarum 118, 64–73.

Groves, M.R., Taylor, M.A., Scott, M., Cummings, N.J., Pickersgill, R.W., Jenkins, J.A., 1996. The prosequence of procaricain forms an alpha-helical domain that prevents access to the substrate-binding cleft. Structure 4, 1193–1203.

Gruis, D.F., Selinger, D.A., Curran, J.M., Jung, R., 2002. Redundant proteolytic mechanisms process seed storage proteins in the absence of seed-type members of the vacuolar processing enzyme family of cysteine proteases. The Plant Cell 14, 2863–2882.

Hayashi, Y., Yamada, K., Shimada, T., Matsushima, R., Nishizawa, N.K., Nishimura, M., Hara-Nishimura, I., 2001. A proteinase-storing body that prepares for cell death or stresses in the epidermal cells of Arabidopsis. Plant Cell Physiology 42, 894–899.

Hensel, L.L., Grbic, V., Baumgarten, D.A., Bleecker, A.B., 1993. Developmental and age-related processes that influence the longevity and senescence of photosynthetic tissues in Arabidopsis. The Plant Cell 5, 553–564.

Holwerda, B.C., Padgett, H.S., Rogers, J.C., 1992. Proaleurain vacuolar targeting is mediated by short contiguous peptide interactions. The Plant Cell 4, 307–318.

Karrer, K.M., Peiffer, S.L., DiTomas, M.E., 1993. Two distinct gene subfamilies within the family of cysteine protease genes. Proceedings of the National Academy of Sciences of the United States of America 90, 3063–3067.

Kervinen, J., Tobin, G.J., Costa, J., Waugh, D.S., Wlodawar, A., Zdanov, A., 1999. Crystal structure of plant aspartic proteinase prophytepsin: inactivation and vacuolar targeting. European Molecular Biology Organization Journal 18, 3947–3955.

Kobayashi, T., Kobayashi, E., Sato, S., Hotta, Y., Miyajima, N., Tanaka, A., Tabata, S., 1994. Characterization of cDNAs induced in meiotic prophase in lily microsporocytes. DNA Research 1, 15–26.

Koizumi, M., Yamaguchi-Shinozaki, K., Tsuji, H., Shinozaki, K., 1993. Structure and expression of two genes that encode distinct drought-inducible cysteine proteinases in *Arabidopsis thaliana*. Gene 129, 175–182.

Kruger, J., Thomas, C.M., Golstein, C., Dixon, M.S., Smoker, M., Tang, S., Mulder, L., Jones, J.D., 2002. A tomato cysteine protease required for *Cf-2*-dependent disease resistance and suppression of autonecrosis. Science 296, 744–747.

Lehfeldt, C., Shirley, A.M., Meyer, K., Ruegger, M.O., Cusumano, J.C., Viitanen, P.V., Strack, D., Chapple, C., 2000. Cloning of the *SNG1* gene of Arabidopsis reveals a role for a serine carboxypeptidase-like protein as an acyltransferase in secondary metabolism. The Plant Cell 12, 1295–1306.

Li, A.X., Steffens, J.C., 2000. An acyltransferase catalyzing the formation of diaclyglucose is a serine carboxypeptidase-like protein. Proceedings of the National Academy of Sciences of the United States of America 97, 6209–6907.

Lu, R., Serrero, G., 2001. Mediation of estrogen mitogenic effect in human breast cancer MSF-7 cells by PC-cell-derived growth factor (PCDGF/granulin precursor). Proceedings of the National Academy of Sciences of the United States of America 98, 142–147.

Mach, L., Mort, J.S., Glossl, J., 1994. Noncovalent complexes between the lysosomal proteinase cathepsin B and its propeptide account for stable, extracellular, high molecular mass forms of the enzyme. Journal of Biological Chemistry 269, 13036–13040.

Mahon, P., Bateman, A., 2000. The PA domain: a protease-associated domain. Protein Science 9, 1930–1934.

McCann, M.C., Stacey, N.J., Roberts, K., 2000. Targeted cell death in xylogenesis. In: Bryant, J.A., Huges, S.G., Garland, J.M. (Eds.), Programmed Cell Death in Animals and Plants. BIOS Scientific Publishers, Oxfordshire, UK, pp. 193–201.

Meichtry, J., Amrhein, N., Schaller, A., 1999. Characterization of the subtilase gene family in tomato (*Lycopersicon esculentum* Mill.). Plant Molecular Biology 39, 749–760.

Murakami, S., Kondo, Y., Nakano, T., 2000. Protease activity of CND41, a chloroplast nucleoid DNA-binding protein, isolated from cultured tobacco cells. FEBS Letters 468, 15–18.

Mutlu, A., Pfeil, J.E., Gal, S., 1998. A probarley lectin processing enzyme purified from *Arabidopsis thaliana* seeds. Phytochemistry 47, 1453–1459.

Nakano, T., Murakami, S., Shoji, T., Yoshida, S., Yamada, Y., Sato, F., 1997. A novel protein with DNA binding activity from tobacco chloroplast nucleoids. The Plant Cell 9, 1673–1682.

Neuteboom, L.W., Veth-Tello, L.M., Clijdesdale, O.R., Hooykaas, P.J., van der Zaal, B.J., 1999. A novel subtilisin-like protease gene from *Arabidopsis thaliana* is expressed at sites of lateral root emergence. DNA Research 6, 9–13.

Pechan, T., Ye, L., Chang, Y., Mitra, A., Lin, L., Davis, F.M., Williams, W.P., Luthe, D.S., 2000. A unique 33-kD cysteine proteinase accumulates in response to larval feeding in maize genotypes resistant to fall armyworm and other Lepidoptera. The Plant Cell 12, 1031–1040.

Ribeiro, A., Akkermans, A.D., van Kammen, A., Bisseling, T., Pawlowski, K., 1995. A nodule-specific gene encoding a subtilisin-like protease is expressed in early stages of actinorhizal nodule development. The Plant Cell 7, 785–794.

Riggs, C.D., Zeman, K., Deguzman, R., Rzepczyk, A., Taylor, A.A., 2001. Antisense inhibition of a tomato meiotic proteinase suggests functional redundancy of proteinases during microsporogenesis. Genome 44, 644–650.

Schluter, U., Muschak, M., Berger, D., Altmann, T., 2003. Photosynthetic performance of an Arabidopsis mutant with elevated stomatal density (sdd1-1) under different light regimes. Journal of Experimental Botany 54, 867–874.

Schoof, H., Zaccaria, P., Gundlach, H., Lemcke, K., Rudd, S., Kolesov, G., Arnold, R., Mewes, H.W., Mayer, K.F., 2002. MIPS *Arabidopsis thaliana* database (MAtDB): an integrated biological knowledge resource based on the first complete plant genome. Nucleic Acids Research 30, 91–93.

Schwechheimer, C., Deng, X.W., 2001. COP9 signalosome revisited: a novel mediator of protein degradation. Trends in Cell Biology 11, 420–426.

Seidah, N.G., Mowla, S.J., Hamelin, J., Mamarbachi, A.M., Benjannet, S., Toure, B.B., Basak, A., Munzer, J.S., Marcinkiewicz, J., Zhong, M., Barale, J.C., Lazure, C., Murphy, R.A., Chretien, M., Marcinkiewicz, M., 1999. Mammalian subtilisin/kexin isozyme SKI-1: a widely expressed proprotein convertase with a unique cleavage specificity and cellular localization. Proceedings of the National Academy of Sciences of the United States of America 96, 1321–1326.

Solomon, M., Belenghi, B., Delledonne, M., Menachem, E., Levine, A., 1999. The involvement of cysteine proteases and protease inhibitor genes in the regulation of programmed cell death in plants. The Plant Cell 11, 431–444.

Steffens, J.C., 2000. Acyltransferases in protease's clothing. The Plant Cell 12, 1253–1256.

Tanaka, H., Onouchi, H., Kondo, M., Hara-Nishimura, I., Nishimura, M., Machida, C., Machida, Y., 2001. A subtilisin-like serine protease is required for epidermal surface formation in *Arabidopsis* embros and juvenile plants. Development 128, 4681–4689.

Tao, K., Stearns, N.A., Dong, J., Wu, Q.L., Sahagian, G.G., 1994. The proregion of cathepsin L is required for proper folding, stability, and ER exit. Archives of Biochemistry and Biophysics 15, 19–27.

Taylor, A.A., Horsch, A., Rzepczyk, A., Hasenkampf, C.A., Riggs, C.D., 1997. Maturation and secretion of a serine proteinase is asso-

ciated with events of late microsporogenesis. The Plant Journal 12, 1261–1271.

Taylor, M.A., Baker, K.C., Briggs, G.S., Connerton, I.F., Cummings, N.J., Pratt, K.A., Revell, D.F., Freedman, R.B., Goodenough, P.W., 1995. Recombinant pro-regions from papain and papaya proteinase IV are selective high affinity inhibitors of the mature papaya enzymes. Protein Engineering 8, 59–62.

Thompson, J.D., Higgins, D.G., Gibson, T.J., 1994. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Research 22, 4673–4680.

Tormakangas, K., Hadlington, J.L., Pimpl, P., Hillmer, S., Brandizzi, F., Teeri, T.H., Denecke, J., 2001. A vacuolar sorting domain may also influence the way in which proteins leave the endoplasmic reticulum. The Plant Cell 13, 2021–2032.

Tornero, P., Conejero, V., Vera, P., 1996. Primary structure and expression of a pathogen-induced protease (PR-P69) in tomato plants: similarity of functional domains to subtilisin-like endoproteases. Proceedings of the National Academy of Sciences of the United States of America 93, 6332–6337.

van den Hazel, H.B., Kielland-Brandt, M.C., Winther, J.R., 1992. Autoactivation of proteinase A initiates activation of yeast vacuolar zymogens. European Journal of Biochemistry 207, 277–283.

Vierstra, R.D., 2003. The ubiquitin/26S proteasome pathway, the complex last chapter in the life of many plant proteins. Trends in Plant Science 8, 135–142.

Visal, S., Taylor, M.A., Michaud, D., 1998. The proregion of papaya proteinase IV inhibits Colorado potato beetle digestive cysteine proteinases. FEBS Letters 434, 401–405.

Von Groll, U., Berger, D., Altmann, T., 2002. The subtilisin-like serine protease SDD1 mediates cell-to-cell signaling during Arabidopsis stomatal development. The Plant Cell 14, 1527–1539.

Wan, L., Xia, Q., Qiu, X., Selvaraj, G., 2002. Early stages of seed development in *Brassica napus*: a seed coat-specific cysteine proteinase associated with programmed cell death of the inner integument. The Plant Journal 30, 1–10.

Yamada, K., Matsushima, R., Nishimura, M., Hara-Nishimura, I., 2001. A slow maturation of a cysteine protease with a granulin domain in the vacuoles of senescing Arabidopsis leaves. Plant Physiology 127, 1626–1634.

Yamagata, H., Masuzawa, T., Nagaoka, Y., Ohnishi, T., Iwasaki, T., 1994. Cucumisin, a serine protease from melon fruits, shares structural homology with subtilisin and is generated from a large precursor. Journal of Biological Chemistry 269, 32725–32731.

Zhao, C., Johnson, B.J., Kositsup, B., Beers, E.P., 2000. Exploiting secondary growth in Arabidopsis. Construction of xylem and bark cDNA libraries and cloning of three xylem endopeptidases. Plant Physiology 123, 1185–1196.