RESEARCH ARTICLE

# *Arabidopsis* cell wall proteome defined using multidimensional protein identification technology

*Emmanuelle M. Bayer, Andrew R. Bottrill, John Walshaw, Marielle Vigouroux, Mike J. Naldrett, Carole L. Thomas and Andrew J. Maule*

John Innes Centre, Norwich Research Park, Norwich, UK

With the completion of the sequencing of the *Arabidopsis* genome and the recent advances in proteomic technology, the identification of proteins from highly complex mixtures is now possible. Rather than using gel electrophoresis and peptide mass fingerprinting, we have used multidimensional protein identification technology (MudPIT) to analyse the 'tightly-bound' proteome for purified cell walls from *Arabidospis* cell suspension cultures. Using bioinformatics for the prediction of signal peptides for targeting to the secretory pathway and for the absence of ER retention signal, 89 proteins were selected as potential extracellular proteins. Only 33% of these were identified in previous proteomic analyses of *Arabidopsis* cell walls. A functional classification revealed that a large proportion of the proteins were enzymes, notably carbohydrate active enzymes, peroxidases and proteases. Comparison of all the published proteomic analyses for the *Arabidopsis* cell wall identified 268 non-redundant genes encoding wall proteins. Sixty of these (22%) were derived from our analysis of tightly-bound wall proteins.

Received: January 31, 2005
Revised: May 17, 2005
Accepted: May 18, 2005

**Keywords:**
*Arabidopsis* / Cell wall proteome / Multidimensional protein identification technology

## 1 Introduction

A major challenge in biology is to identify the complete proteomes of subcellular components and compartments. Plant cells are surrounded by a supportive cellulosic wall that shapes the cell and provides an environment for extracellular interactions. Although superficially rigid, the wall is an extremely dynamic structure with roles in cell growth and development, signalling, plant defence, and intercellular communication [1]. The carbohydrate component of the wall accounts for about 90% of the dry weight and constitutes a framework of cellulose microfibrils embedded in a matrix of hemicellulose and pectins [2]. In contrast, proteins contribute less than 10% to the wall mass; this may nevertheless represent several hundred proteins [1, 3], which are essential for maintaining and regulating the physical and biological functions of the plant extracellular matrix.

The identification of the complete cell wall proteome presents several significant technical hurdles [1]: (1) Plant tissues are complex mixtures of cell types with differing cell wall characteristics. (2) Preparations of plant cell walls do not comprise particulate fractions of uniform size and density and are therefore difficult to obtain in 'pure' form. (3) Many proteins are tightly associated with the wall and require strongly chaotropic and denaturing conditions for their isolation and characterisation. (4) Isolated walls contain some hydrophobic and many basic proteins [3, 4], neither class of which are well resolved using 2-DE.

In recent years, various technical approaches have been followed to establish the wall proteome for *Arabidopsis thaliana* [3–7]. In all cases, gel electrophoresis and PMF have been used for protein identification. Ionic solutions have been used to elute loosely-bound wall proteins from intact tissues [4, 6] or suspension cells [3, 5]. In one case, Chivasa *et*

**Correspondence:** Dr. A. J. Maule, John Innes Centre, Norwich Research Park, Colney, Norwich NR4 7UH, UK
**E-mail:** andy.maule@bbsrc.ac.uk
**Fax:** +44-1603-450-045

**Abbreviations: GPI**, glycosylphosphatidyl inositol; **HRGP**, hydroxyproline-rich glycoprotein; **LRR**, leucine-rich repeat; **LRX**, leucine-rich repeat extensin; **MudPIT**, multidimensional protein identification technology; **XTH**, xyloglucan endotransglucosylase-hydrolase

*al.* [5] performed a proteomic analysis, using 2-DE/MALDI-TOF-MS, of proteins sequentially extracted from enriched *Arabidopsis* cell wall fractions. The proteins were isolated using a two part extraction, a mild extraction buffer containing $CaCl_2$, and a stronger urea-containing buffer. In a complementary approach, we have used multi-dimensional protein identification technology (MudPIT) to identify tightly-bound cell wall proteins extracted from purified wall preparations isolated from *Arabidopsis* suspension cells using strong chaotropic and denaturing reagents. Unlike 2-DE approaches, which bias against membrane-associated proteins, less abundant proteins and proteins with extreme p*I* or molecular weight (MW), MudPIT is a relatively unbiased technology allowing the identification of different classes of protein with a similar sensitivity. The technology couples 2-D LC to MS/MS [8–10]. Peptides generated enzymatically and/or chemically from a complex protein mixture are separated by the 2-D LC, based upon their charge and hydrophobicity, and then sequenced by MS/MS. The multidimensional chromatography achieves a high-resolution separation of peptides facilitating the downstream MS. When compared with gel electrophoresis-based methods, MudPIT analysis generally identifies a greater number of proteins [8]. For instance, Washburn *et al.* [10] successfully identified a total of 1484 proteins from yeast cell lysates and an analysis of proteins extracted from rice yielded 2363 unique proteins [11]. In our experiments, MudPIT identified a total of 792 unique protein sequences. Using bioinformatic softwares for subcellular prediction, 89 sequences were selected as a subset of extracellular proteins. This work substantially extends our knowledge of the cell wall proteome, particularly for the more tightly-bound classes of proteins.

## 2 Materials and methods

### 2.1 Cell wall preparation and purity assessment

The growth of *A. thaliana* (ecotype Landsberg erecta) suspension cultures, the preparation of cell wall fractions and their purity assessment are described in Bayer *et al.* [12]. In brief, 3-day-old suspension culture cells were collected and resuspended with cold wall-preparation buffer (100 mM Tris-HCl, pH 8.0, 100 mM KCl, 10% (v/v) glycerol, 10 mM EDTA, 0.45 M mannitol and a complete protease inhibitor cocktail (Roche Molecular Biochemicals)). Cells were processed through an $N_2$ disruption bomb and the recovered walls were further ground in liquid $N_2$ with a mortar and a pestle. The frozen powder was thawed and the walls washed extensively in cold, wall-washing buffer (100 mM Tris-HCl, pH 8.0, 100 mM NaCl, 10% (v/v) glycerol, 10 mM EDTA). Pelleted walls were kept at −20°C. Before protein extraction, purified wall fragments were extracted for 30 min, with shaking at room temperature with 200 mM $CaCl_2$ (0.8 mL/10 mL starting culture). The walls were recovered by centrifugation at 10 000 × g for 30 min and washed three times in water.

### 2.2 Protein extraction

Pelleted, washed walls from 500 mL of starting culture were further extracted with 1 volume of detergent/reducing buffer (1.8% SDS, 2% DTT, HEPES.KOH 100 mM pH 7.5, 5% glycerol, 0.5% polyvinylpyrrolidone, 50 mM sodium pyrophosphate) for 2 h at room temperature and a further 30 min at 50°C. The protein extract was separated from the walls by centrifuging for 30 min, at 20 000 × g, at 25°C then mixed with 1.4 volumes of Tris-buffered phenol pH 8. After centrifugation, the protein sample was back-extracted four times with 100 mM Tris-HCl, pH 8.4, 20 mM KCl, 10 mM EDTA and 0.4% β-mercaptoethanol. The upper phases were kept after each back extraction, pooled and proteins precipitated with 80% cold acetone. The proteins were precipitated from the phenol phase using 5 volumes of 100 mM ammonium acetate in methanol at −20°C. The precipitated proteins were then recovered by centrifugation at 4°C, 30 min at 10 000 × g and washed twice with 100 mM ammonium acetate in methanol and four times with 80% acetone. To assess the complexity of the protein mixtures, the preparations were analysed using 2-D electrophoresis as described in Peck *et al.* [13]. Colloidal CB staining was performed as described in Neuhoff *et al.* [14] and gels destained in water. The gels were then scanned using a Bio-Rad GS-710 calibrated imaging densitometer.

### 2.3 Protein identification by MudPIT

Proteins were extracted from 15 mL of pelleted, $CaCl_2$–washed walls with the detergent/reducing buffer (1.8% SDS, 2% DTT, HEPES.KOH 100 mM pH 7.5, 5% glycerol, 0.5% polyvinylpyrrolidone, 50 mM sodium pyrophosphate). These proteins were then extracted with Tris-buffered phenol pH 8 and, precipitated with ammonium acetate in methanol as described in the Sect. 2.2. Approximately 1 mg of dry pelleted protein was recovered from 15 mL of walls.

For protocol 1, the entire protein pellet was resuspended in 400 μL of urea solubilisation buffer (9 M urea, 0.5% DTT and 0.5% IPG-buffer pH 3–10), then washed twice with the same buffer. After centrifugation, the remaining pellet was dissolved in 50 μL of Laemmli buffer (62.5 mM Tris-HCl pH 6.8, 0.5% (w/v) SDS, 4.5% (v/v) β-mercaptoethanol, 10% (v/v) glycerol and 0.025% (w/v) bromophenol blue). The proteins solubilised in the 400 μL of urea buffer were precipitated overnight, at −20°C from 95% cold acetone. The proteins were recovered by centrifugation at 10 000 × g for 30 min at 4°C, washed twice with 80% acetone and air dried before being resuspended with 50 μL of Laemmli buffer. The two protein samples were loaded on 10% acrylamide gels for SDS-PAGE analysis but the migration was stopped just as the proteins were entering the separating gel so that they were concentrated on a very fine band. This step removed non-proteinaceous contaminants from the sample. After colloidal CB staining, the bands were excised from the gel, washed, reduced with DTT, *S*-alkylated with iodoacetamide,

 

then in-gel digested with trypsin (sequencing grade modified trypsin; Promega, Madison, WI, USA) using an Investigator™ ProGest™ protein digestion station (Genomic Solutions, Huntingdon, UK).

For protocol 2, the protein pellet was first redissolved and chemically cleaved by treatment for 40 h in 500 μL of 70% formic acid containing 45 mg of cyanogen bromide; the high concentration of formic acid was necessary to assist solubilisation. The sample was lyophilised after addition of 500 μL of 10 mM Tris-HCl pH 8.0. A proportion of the cleaved proteins (200 μg) was resuspended in 8 M urea, 50 mM Tris-HCl pH 8.0 and reduced and alkylated with Tris (2-carboxyethyl) phosphine and iodoacetamide, respectively. The sample was diluted to 4 M urea and digested with endoproteinase Lys-C (Roche Applied Science) for 8 h at 37°C, diluted to 2 M urea and further digested with trypsin for 16 h. To prepare the sample for 2-D LC, procedures similar to those in [10] and [9] were followed, but use of ammonium bicarbonate was avoided so as to allow direct loading onto the strong cation exchange column without prior desalting.

A 75-μm PicoFrit capillary (New Objective, Town Country) was packed first with 90 mm of Symmetry™ C18 300 Å reverse-phase material (Waters, Elstree, UK), followed by 30 mm of PartiSphere™ strong cation exchange material (Whatman International, Maidstone, UK), using an in-house high pressure packing device, dimensions of which were kindly provided to us by Dr. John Yates III, Scripps Research Institute, CA, USA. The resulting biphasic microcapillary column was equilibrated to 5% ACN/0.1% formic acid. After loading the sample, the column was mounted on the Z-Spray™ ion source of a Q-TOF2™ mass spectrometer (Micromass, Manchester, UK), and in-line with a capillary HP LC system (CapLC, Waters). A fully automated multistep chromatography run was carried out, with the mass spectrometer operating in data-dependent mode during each RP elution. The three buffer solutions used for chromatography were 0.1% formic acid (buffer A), 100% ACN/0.1% formic acid (buffer B) and 500 mM ammonium acetate/0.1% formic acid (buffer C). The peptides were first separated on the ion exchange column using a series of gradient elutions between increasing salt concentrations (from x% to y% over 15 min and then held at y% for 15 min, where x and y were consecutive values in the following series: For protocol 1: 0, 20, 50, 75, 100, 150, 200, 250, 375 and 500 mM; for protocol 2: 0, 10, 20, 30, 40, 50, 60, 75, 100, 200, 300 and 500 mM. The flow rate was maintained at 0.5 μL/min during the salt gradients). Each salt gradient was followed by RP separation for which an increasing concentration gradient of buffer B was used to elute the peptides (from 5 to 40% over 80 min then from 40 to 80% over 10 min). The overall duty cycle time for the MudPIT analysis was 20 h for protocol 1 and 24 h for protocol 2. Mass spectra were acquired as follows: After a 1.2-s MS scan, the three most intense signals were subjected to MS/MS, with scan times of 1.2 s each, before returning to MS mode. Ion spectra were searched using an in-house version of the MASCOT search tool (Matrix Science, London, UK, http://www.matrixscience.com; [15]) against a bi-monthly updated copy of the UniProt database. The taxonomy was limited to *A. thaliana*. The sole fixed modification was carboxyamidomethyl-CYS and the only variable modification was oxidation of MET. The enzyme was selected as 'trypsin' (protocol 1) or 'trypsin and CNBr' (protocol 2) and the maximum number of missed cleavages was set as 3. MS/MS fragment ion tolerances were set at ±0.25 Da. The criteria for protein identification were based on the manufacturer's definitions (Matrix Science). Candidate peptides with probability-based Mowse scores exceeding threshold ($p < 0.05$), and thus indicating significant or extensive homology, were referred to as 'hits'. Data from protocol 1 were scored using 'Standard' scoring. Data from protocol 2 were scored using 'MudPIT' scoring, as defined by Matrix Science.

### 2.4 Bioinformatic analysis

The presence of signal peptides and transmembrane domains was predicted using the Signal-P V3 programme (http://www.cbs.dtu.dk/services/TargetP/; [16]) and TMHMM2 (http://www.cbs.dtu.dk/services/TMHMM; [17, [18]), respectively. Glycosylphosphatidyl inositol (GPI)-anchored proteins were determined using two prediction tools, DGPI (http://www.expasy.org/tools/) and Big-PI plant predictor (http://mendel.imp.univie.ac.at/gpi/gpi_server.html; [19]). The presence of conserved domains in proteins predicted to contain a signal peptide was searched for using Pfam (http://www.sanger.ac.uk/Software/Pfam/search.shtml). The carbohydrate active enzymes were classified into families according to the CAZY nomenclature (http://afmb.cnrs-mrs.fr/CAZY/).

### 2.5 Comparison of identified proteins with those from other proteomics studies

In order to compare proteins between two datasets, each member of one set was compared with each of the other. The first comparison was at the level of accession codes. For the remainder, they were compared using the NEEDLE global alignment software within the EMBOSS package [20]. The threshold for screening possible matches of interest was set at 95% when our data were compared with previous cell wall or plasma membrane proteomes, and 99% when the different MudPIT lists generated in this work were compared against each other. The procedure was automated with a Perl script. A cut-off of 95% was selected as accessions between different protein databases may include small sequence variations. A cut-off of 99% identity within experiments, rather than 100%, served to accommodate the fact that the MudPIT approach cannot always distinguish between highly homologous proteins. As well as considering the percentage identity of the entire alignment, for each sequence a second identity score was parsed from the output, namely, the pro-

 

portion of a whole sequence, which was identical to the corresponding part of the other sequence. This was to identify situations where two matching sequences represented the whole and a fragment of the same protein.

The establishment of *A. thaliana* cell wall proteome was done by comparison of the five proteomic studies of the *Arabidopsis* cell wall (present study and [3–6]). To that end, the Uniprot accession numbers for the 89 cell wall proteins were converted into their respective gene identification (AGI) numbers. Redundant identical AGI numbers within the complete list were discarded.

## 3 Results and discussion

### 3.1 Purified cell walls

For this work, a rigorous purification procedure was applied to obtain a clean cell wall fraction from *Arabidopsis* suspension cells. This fraction, containing exclusively primary cell walls, was shown by transmission electron microscopy and immunoblot analysis [12] to be virtually free of cytoplasmic contaminants and plasma membrane although, occasionally, small blebs of membrane were seen to remain associated with the walls. We have estimated [12] that there was a several thousandfold reduction in contaminant proteins when compared to crude cell extracts. Nevertheless, using highly purified cell walls as a starting material for proteomics is particularly important for MudPIT as the technique provides a relatively unbiased and mostly qualitative output that could include many of the contaminant proteins. The presence of membrane on the wall fragments was expected, as during plasmolysis the contracting protoplast remains connected with the cell wall by threads of plasma membrane, referred to as Hechtian strands. The origin of these strands at plasmodesmata and the plasma membrane-wall adhesion sites may cause small portions of plasma membrane to remain associated with the purified walls. Plasmolysis was an integral step of the purification protocol. Our wall purification procedure included multiple washes in ionic buffers (NaCl) and EDTA, and a final wash in $CaCl_2$ to release the loosely-bound proteins.

### 3.2 Cell wall protein extraction and protein identification using MudPIT

A two-phase buffered phenol extraction was routinely performed directly on the cell wall protein extract in order to remove pectin from the sample and to prevent the formation of a gel-like solution. Pectin represents about 30% dry weight of the primary plant wall and is the most soluble of the wall polysaccharides [2]. During the two-phase buffered phenol isolation, the proteins were partitioned into the phenol phase and the polysaccharides into the aqueous phase. Successive back-extraction steps with buffer ensured the total removal of the pectin without protein loss into the aqueous phase (data not shown).

In preparation for the MudPIT analysis, we reasoned that the potential for maximal protein identification would increase if we could reduce the complexity of the peptide mixtures in a single sample. To achieve this (Protocol 1), the protein extract from 15 mL of cell wall preparation was divided based upon differential solubility in a urea buffer lacking detergent and in a strongly solubilising buffer including thiourea, CHAPS and a high concentration of DTT. For a provisional comparison of the relatively complexity of these two samples, they were directly analysed by 2-DE. Each of the solubilisation steps showed a distinct pattern of protein spots (data not shown). For MudPIT analysis, the proteins in each sample were captured within a small gel piece and in-gel digestion carried out using trypsin. MudPIT identified 416 and 70 proteins, for the urea and detergent/reducing solubilisation buffer extractions, respectively (supplementary data lists 1 and 2). When these two protein lists were compared, only approximately half of the proteins in the second sample were unique, suggesting that the differential extraction to reduce complexity may have been unnecessary. Between the two samples, 31 common proteins were found, leaving 455 unique protein sequences.

In MudPIT Protocol 2 (see Sect. 2), the extracted proteins were cleaved directly using cyanogen bromide. This was followed by enzymatic digestion in 4 M urea with Lys-C and in 2 M urea with trypsin. In this case, the combination of chemical and enzymatic cleavage aimed to increase the number of peptides. Lys-C is active in 4 M urea and cleaves after lysine while trypsin is active at lower urea concentrations and cleaves after lysine or arginine. The sequential enzyme digestion was also intended to increase peptide frequency, achieved also after the exposure of more target sites in higher concentrations of urea. With larger numbers of peptides, the identification of individual peptides and peptide coverage for individual proteins should have been maximised.

The second protocol led to the identification of 556 unique protein sequences (Supplementary Data list 3). All the identifications had scores (calculated by MASCOT programme) greater or equal to the significance threshold, determined at the 95% confidence level. In the majority of cases, these were based upon multiple peptide matches. Access to the complete sequence of the *Arabidopsis* genome sequence meant, however, that even single peptide sequences could be used in gene identification. Occasionally, however, single or a few peptides identified gene families. Comparison of the three MudPIT lists revealed 792 unique protein sequences (Supplementary Data list 4).

### 3.3 Identification of classical cell wall proteins

Because of the high sensitivity and largely qualitative output of MudPIT, we expected the MudPIT lists to contain *bona fide* cell wall proteins and contaminant proteins. Notwithstanding the purity of the cell wall preparation [12], contaminant proteins were to be expected, particularly con-

 

sidering the large amount of cell wall material used for the MudPIT analyses. In addition, given the acidic nature of the extracellular matrix, there may be particular opportunities for cytoplasmic proteins to bind to walls through ionic interactions. It is also possible that some of the proteins predicted to be contaminants could have arrived in the cell wall via an alternative pathway. Proteins that do not contain a secretory signal peptide have been isolated from purified cell walls in previous studies and the existence of an alternative secretory pathway that might not require transition through the endomembrane system has been suggested [5, 21, 22].

Classical cell wall proteins usually contain a cleavable signal peptide on the N terminus of the protein precursor, which is responsible for their targeting to the ER. The proteins destined for secretion leave the ER for the Golgi complex where they are assembled within vesicles, and subsequently released via exocytosis, in the extracellular matrix. As a result classical cell wall proteins do not contain the KDEL or HDEL ER retention motifs, found at the C terminus of the protein. Borderies *et al*. [3], working with extracellular proteins from *Arabidopsis* suspension cells, used these criteria for distinguishing cell wall proteins from contaminants. To identify signal peptides for secretion we used the Signal P V 3 software. These signal peptides do not have a consensus amino acid sequence but are characterised by different conserved domains (positively charged, hydrophobic, polar). SignalP contains two programmes: SignalP-NN and SignalP-HMM both predicting the presence or absence of a signal sequence as well as signal peptidase cleavage sites. Because SignalP-NN was shown to be slightly better in predicting signal peptidase cleavage sites [16], only this prediction was used. Importantly, not all signal peptides are cleaved, and in some cases they may serve to anchor the protein into the ER phospholipid bilayers. SignalP-HMM was used to distinguish between cleaved and uncleaved signal peptides. Of the total list of protein sequences generated by the MudPIT analysis, 116 proteins were found to contain a signal peptide. Proteins were selected as potential wall-associated proteins if they contained a signal peptide and if they lacked an ER retention motif. Of the 116 proteins predicted to be translocated into the ER compartment, only 10 (Q9LKR3, Q39043, O04151, O48773, Q8H1B3, Q8RVG8, Q8LAM5, Q8LC60, Q8LC80, Q8H792) were found to contain an HDEL or KDEL ER retention signal. Some of these proteins (*e.g.* binding proteins; Q9LKR3, Q39043) are well-known ER residents, are relatively abundant in this compartment, and have been detected before in our wall extracts (*e.g.* BiP; [12]). Since we expect small portions of the cortical ER to co-purify with the wall, *i.e.* as a component of, or associated with plasmodesmata [12], the detection of some ER proteins was expected. Nevertheless, the number of ER proteins remains very low. On this basis, 106 potential cell wall proteins (*i.e.* proteins containing a signal peptide but no ER retention signal) were selected.

From our MudPIT list, 15% of the wall-extracted proteins were predicted to contain a signal peptide. This percentage is lower than reported in related work [3, 5, 21], although direct comparisons may be difficult owing to the different systems and technologies used. MudPIT has the potential to resolve a very large number of proteins and low abundance wall-associated proteins can be more easily detected. However, this also holds true for contaminant proteins.

### 3.4 Membrane proteins

As discussed above we were not surprised to find membrane proteins in our cell wall extracts. Although, proteins containing transmembrane domains are generally not considered as classical cell wall proteins [3, 5], our observations that purified cell walls retain membranous components indicated that these should be included in this study. Integral plasma membrane proteins also possess a signal peptide and no ER retention motif. To allow their anchoring into the phospholipid bilayer, they contain an unusually long stretch of hydrophobic residues, which form transmembrane helices. Several classes of proteins, such as the wall-associated protein kinases (WAK) are known to be associated both with the plasma membrane and the extracellular matrix, and yet form structural and functional elements of the plasma-membrane-cell wall continuum, playing a vital role for the transmission of signals [23]. Of the 106 proteins found to contain a signal peptide but no ER retention signal, the signal sequence was predicted to serve as a membrane anchor for only 5 proteins (Q94EJ6, Q9C8H5, Q8VZV7, Q9C568 and P92994).

In order to distinguish between "free" and membrane-associated extracellular proteins, we used the TMHMM 2.0 software [17, 18] to predict the presence of transmembrane helices. Although the prediction accuracy has been estimated to be around 83% [18, 24], a frequent source of false-positive results are signal peptides that target the protein for export. Indeed, the targeting sequence contains a hydrophobic region that is easily mistaken for a transmembrane region by most prediction programmes [24]. TMHMM 2.0 was found to erroneously identify around 20% of the signal sequence as transmembrane helices [18, 24]. Therefore, if the predicted transmembrane helix coincided with a predicted cleavable signal peptide, the transmembrane domain was not taken into consideration. Of the 106 proteins, 20 were predicted to contain one or several transmembrane domains. A third of those proteins are predicted to be plasma membrane kinase-like proteins, all containing an extracellular leucine-rich domain (LRR) for protein interaction.

### 3.5 Post-translational modification

Some extracellular proteins are known to interact with the plasma membrane via a lipid anchor, such as GPI anchor. All known GPI-anchored proteins possess, in addition to the cleavable N-terminal signal peptide for secretion, a signal

 

sequence for GPI-anchoring that lies at the very C terminus of the polypeptide chain (30–40 residues). This GPI signal consists of four domains: a polar flexible region, a region of small residues containing a cleavage site, a hydrophilic spacer domain and a hydrophobic domain, which is thought to form a transient transmembrane domain [25]. The phosphatidylinositol-moiety of the GPI anchor is susceptible to cleavage by specific phospholipases, and consequently GPI-anchored proteins exist in both a soluble and a membrane associated form.

Using two GPI-anchored protein prediction tools, DGPI and Big-PI plant predictor, 18 and 3 proteins, respectively, were predicted to contain a GPI anchor. Three proteins were identified by both programmes (Q9FFH6, Q9FPD6, Q9FHX5). For 5 of the 18 predicted GPI-anchored proteins (Q94CD8, Q9FFH6, Q9FHX5, Q9FNQ2 and Q9LW07) the presence of a GPI-moiety had been confirmed experimentally [26, 27]. No cleavage sites could be detected for six of the putative GPI-anchored proteins (Q9FF10, Q8GXV7, O22940, Q8LDT6, Q9M1J5, Q9ZPS7), which raised doubts about the significance of the prediction. Three of the six proteins (Q8GXV7, Q8LDT6, Q9ZPS7) may nevertheless be associated with a phospholipid bilayer as transmembrane domains were predicted by the TMHMM2 programme. The Big-PI predictor was specifically designed to predict GPI lipid anchoring of plant proteins and according to the authors false predictions represent only 0.1% [19]. Because DGPI predicted some non-extracellular proteins to be GPI-anchored (for instance Q41951), the overall reliability of the method may be questioned. However, three of the proteins positively predicted by DGPI but not Big-PI (Q9FNQ2, Q94CD8 and Q9LW07), have been shown experimentally to be GPI anchored [26]. Only further experiments would determine if the other proteins are indeed associated with the plasma membrane by a GPI-moiety.

Recent proteomic and genomic analysis in *Arabidopsis* and rice established that most of the predicted GPI-anchored proteins belong to functional families for cell wall synthesis and remodelling, including glycohydrolases of family 17 (β-1,3 glucanases), extensin-like proteins, fasciclin-like proteins and COBRA-like proteins [27, 28]. Extracellular arabinogalactan-rich proteins (AGP) are also found anchored to the plasma membrane via a GPI anchor. Nevertheless, the types of plant proteins that use this mode of attachment are largely unknown, with 20–25% classified as hypothetical proteins [19]. Among the GPI-anchored proteins identified in this study, the glycosyl hydrolase family is the best represented class (including three β-1,3 glucanases and two β-1,4 glucanases).

### 3.6 Classification of cell wall proteins

The potential cell wall proteins (containing a signal peptide but no ER retention motif; see Table 1 and Supplementary Data 5 for details) were classified into functional classes by sequence homology with known proteins or identification of functional domains. Putative functions could be assigned for 85% of the proteins. Primary cell walls are known to contain numerous enzymes [3, 29] and it is therefore not surprising that the majority of the proteins fell into this class (45%). Cell wall enzymes are all hydrolases, or oxido-reductases that operate with simple substrates (*e.g.* $H_2O$, $H_2O_2$, $O_2$) found in the extracellular matrix.

The first category of enzyme is composed of the carbohydrate active enzymes (CAZymes). The CAZyme database so far includes only four main classes; the glycoside hydrolases, glycosyltransferases, polysaccharide lyases and the carbohydrate esterases (http://afmb.cnrs-mrs.fr/CAZY/). In our work, the glycoside hydrolase group was certainly the most represented class of enzymes, covering 9 of the 32 families in *Arabidopsis* [30], some of them containing several members. The endo-1,3-β glucanases (family 17) represented the largest group with four members, three of which were predicted to be membrane-associated via a GPI-anchor. In plants, the endo-1,3-β glucanases are found in the extracellular matrix or in the vacuole. In the former case, the enzymes are generally acidic whereas in the latter, the basic form predominates [31]. One of the four endo-1,3-β glucanases (Q8LB06) was found to have a predicted basic p*I* (around 8.7), suggesting a vacuolar location. However, the protein was also identified as a wall protein by Chivasa *et al.* [5]. Some of the glycoside hydrolases such as xyloglucan endotransglucosylase-hydrolases (XTH, family 16) and endo-1,4-β-D-glucanases (family 9) may act as wall loosening agents for cell wall elongation [32] by cleaving the xyloglucans that link cellulose microfibrils together. XTH have also been proposed to ligate the newly synthesised xyloglucans secreted into the extracellular matrix by catalysing the intramolecular cleavage of a xyloglucan polymer and transferring the newly generated end to a new xyloglucan polymer [33].

Two other classes of hydrolases were identified: two members of the carbohydrate esterase class (family 8 and 13), and nine members of the protease family. The protein sequence Q8L9B9 was found to be annotated in UniProt as a putative chloroplast nucleoid DNA binding protein. However, analysis of the sequence with Pfam revealed the presence of a domain characteristic of eukaryotic aspartyl protease. It seems very likely that Q8L9B9 has been mis-annotated as chloroplastic DNA binding protein instead of apoplastic protease. Like the endo-1,3-β glucanases, the cell wall proteases are thought to play a role in cell defence against pathogen attacks. As mentioned by Borderies *et al.* [3], defence proteins may be expected when working with cell cultures because of the stress associated with the culture conditions. Although in their study, defence proteins represented the second most abundant group of proteins, no endo-1,3-β glucanases and only one protease (Q9SVD1) was found in common between the two studies.

Expansins have been proposed to regulate cell growth by breaking hydrogen bonds between cellulose and xyloglucan [34]. Two classes of expansins are currently recognised, the α-expansin and β-expansin. The *Arabidopsis* genome contains

 

**Table 1.** List of *Arabidopsis* cell wall proteins[a)]

| Family[b)] | UniProt Accession number[c)] | Locus (AGI) | Description | MW (kDa) | p*I* |
|---|---|---|---|---|---|
| Glycosyl hydrolase family 1 | Q9LV33* | At3g18080 | Putative beta-glucosidase | 59.1 | 9.44 |
| | O64879* | At2g44450 | Putative beta-glucosidase | 52.2 | 7.56 |
| Glycosyl hydrolase family 3 | Q8W112* | At5g20950 | Beta-D-glucan exohydrolase-like protein | 68.4 | 9.23 |
| | Q8VZG5 | At1g78060 | Putative protein beta xylosidase | 84.7 | 8.3 |
| Glycosyl hydrolase family 9 | Q8LDE8 | At1g71380 | Putative beta-glucanase | 53.6 | 8.81 |
| | O23134* | At1g22880 | putative endo-1,4-beta-D-glucanase | 54 | 9.3 |
| Glycosyl hydrolase family 16 | Q39099* | At2g06850 | AtXTH 4 | 34.5 | 9.03 |
| | O80803 | At1g65310 | AtXTH 17 | 32.3 | 8.79 |
| | P24806 | At4g30270 | AtXTH 24 | 30.8 | 8.35 |
| Glycosyl hydrolase family 17 | Q9FNQ2 | At5g61130 | Putative glycosyl hydrolase family 17 protein | 20.9 | 7.46 |
| | Q9FHX5* | At5g42100 | Putative glucan endo-1–3-beta-glucosidase | 45.6 | 7.03 |
| | Q8LB06* | At3g07320 | Putative glucan endo-1–3-beta-glucosidase | 50.9 | 8.67 |
| | Q94CD8 | At3g13560 | Putative glucan endo-1,3-beta-glucosidase | 54.8 | 5.9 |
| Glycosyl hydrolase family 27 | Q8RWB9 | At5g08380 | Alpha-galactosidase-like protein | 46 | 6.32 |
| | Q8RX86 | At5g08370 | Alpha-galactosidase-like protein | 44.4 | 7.55 |
| Glycosyl hydrolase family 28 | Q9LW07 | At3g15720 | Polygalacturonase-like protein | 55 | 5.36 |
| Glycosyl hydrolase family 31 | Q9S7Y7 | At1g68560 | Alpha-xylosidase | 102.9 | 6.31 |
| Glycosyl hydrolase family 79 | Q9FF10 | At5g07830 | Similarity to heparanase | 60.8 | 9.08 |
| Carbohydrate esterase family 8 | Q42534 | At1g53830 | Pectin methylesterase 2 | 64.6 | 9.08 |
| Carbohydrate esterase family 13 | Q940J8* | At4g19410 | Putative pectin acetylesterase | 43 | 9.31 |
| Proteases | Q9FPD6 | At1g08210 | Putative protease protein (aspartyl protease) | 53.7 | 5.73 |
| | Q9MAP5 | At1g32960 | putative subtilase (serine protease) | 84.2 | 6.27 |
| | Q9SVD1 | At3g52500 | Hypothetical protein (aspartyl protease) | 51.7 | 8.53 |
| | Q9C6Y5 | At1g44130 | Putative nucellin (aspartyl protease) | 44.6 | 8.87 |
| | Q940R4 | At4g16563 | Unknown protein (aspartyl protease) | 55.5 | 8.61 |
| | Q9M356 | At3g61820 | Hypothetical protein (aspartyl protease) | 52.1 | 8.76 |
| | Q8L9B9 | At1g01300 | Putative aspartyl protease | 53 | 9.4 |
| | Q9C7E2* | At1g28110 | Putative serine carboxypeptidase II | 51.6 | 6.86 |
| | Q8VY01 | At2g33530 | Putative serine carboxypeptidase II | 52.2 | 8.66 |
| Protease inhibitors | Q39182 | At2g02100 | Low-molecular-weight cysteine-rich protein | 8.9 | 9.37 |
| Peroxidases | Q96506 | At1g05240 | Peroxidase 1/2 | 36 | 9.35 |
| | Q96520* | At1g71695 | Peroxidase 12 | 39.9 | 8.58 |
| | Q9LSY7* | At3g21770 | Peroxidase 30 | 36.2 | 9.71 |
| | Q9LHA7 | At3g28200 | Peroxidase 31 | 35.7 | 9.22 |
| | Q43729 | At5g17820 | Peroxidase 57 | 34.5 | 10.1 |
| | Q43872 | At5g42180 | Peroxidase 64 | 35.1 | 9.05 |
| Other enzymes | Q9ZWC4* | At1g04040 | Similar to acid phosphatase | 31.4 | 9.22 |
| | Q9SIV9* | At2g16430 | Putative purple acid phosphatase | 54.6 | 7.3 |
| | Q94EJ6 | At4g18030 | Putative methyl transferase (DUF248) | 71.2 | 7.79 |
| | Q9SVE3 | At4g38420 | Putative pectinesterase | 62.6 | 9.59 |
| | Q9M263 | At3g62020 | Germin-like protein | 23.7 | 8.94 |
| | Q9SVG4 | At4g20830 | Reticuline oxidase-like protein | 60.4 | 9.55 |
| | P92994 | At2g30490 | Trans-cinnamate 4-monooxygenase cytochrome P450 family | 57.9 | 8.89 |
| | P92976* | At1g74000 | Strictosidine synthase 3 | 34.8 | 9.66 |
| | Q9C9C2 | At1g74010 | Putative strictosidine synthase | 34.3 | 8.26 |

 

**Table 1.** Continued

| Family[b)] | UniProt Accession number[c)] | Locus (AGI) | Description | MW (kDa) | p*I* |
|---|---|---|---|---|---|
| Expansins | Q9LZT4 | At3g45970 | Expansin-like 1 family A | 29.2 | 8.28 |
| | Q9LZT5 | At3g45960 | Expansin-like 3 family A | 29.1 | 9.2 |
| | Q9C554 | At1g69530 | Alpha-expansin 1 | 26.9 | 9.26 |
| | O48818 | At2g39700 | Alpha-expansin 4 | 28.3 | 9.76 |
| | Q38865 | At2g28950 | Alpha-expansin 6 | 28.2 | 9.76 |
| LRR-Extensins | O48809 | At1g62440 | AtLRX2 | 86.2 | 5.88 |
| | Q9T0K5* | At4g13340 | AtLRX3 | 83 | 6.49 |
| | Q9LHF1* | At3g24480 | AtLRX4 | 55.3 | 6.46 |
| Lectin | Q8H788 | At1g78830 | Hypothetical protein (lectin motif) | 50.6 | 8.66 |
| | Q94K76 | At5g18470 | Hypothetical protein (lectin motif) | 46.2 | 6.45 |
| LRR | Q9M5J9 | At5g06860 | Polygalacturonase-inhibiting protein 1 | 37.3 | 8.24 |
| | Q9M5J8* | At5g06870 | Polygalacturonase-inhibiting protein 2 | 37.5 | 9.02 |
| | Q8LGC3 | At1g33590 | Putative disease resistance protein | 52.3 | 9.23 |
| | O48809 | At1g62440 | Hypothetical protein | 86.2 | 5.88 |
| LRR Kinases | Q9FMD7 | At5g16590 | Receptor-like protein kinase | 68 | 8.76 |
| | Q9SIT1 | At2g01820 | Putative receptor-like protein kinase | 102.6 | 5.93 |
| | Q9C9Y8 | At3g08680 | Hypothetical protein. | 69.9 | 7.62 |
| | Q9SCT4* | At3g51740 | Hypothetical protein | 91 | 7.62 |
| | Q9M8T0* | At3g02880 | Hypothetical protein | 68.1 | 8.6 |
| | Q9SIX4 | At2g16250 | Putative LRR receptor protein kinase | 101.8 | 8.5 |
| | Q9ZU46 | At2g01210 | Putative receptor-like protein kinase | 79.3 | 5.75 |
| Phosphate induced | Q8LB34* | At4g08950 | Phi-1-like protein | 33.9 | 9.39 |
| | Q9FE06* | At5g64260 | Phi-1-like protein | 33 | 9.45 |
| | Q9C6E4 | At1g35140 | Putative Phi-1-like protein | 33.5 | 8.93 |
| Unknown/Hypothetical proteins | Q9FFH6 | At5g44130 | Hypothetical protein (fasciclin domain) | 26.2 | 5.53 |
| | Q9LQ75 | At1g01800 | Unknown protein | 35.4 | 5.29 |
| | Q8VZV7 | At5g14430 | Hypothetical protein (DUF 248) | 70.6 | 9.06 |
| | Q8GXV7 | At5g43980 | Hypothetical protein (DUF26) | 32.7 | 5.51 |
| | O22940* | At2g41800 | Unknown protein (DUF642) | 40.6 | 9.22 |
| | Q8LEX7 | At5g25460 | Hypothetical protein (DUF 642) | 40.2 | 6.92 |
| | Q9C8H5 | At1g51630 | Hypothetical protein (DUF 246) | 51.9 | 6.68 |
| | Q8LBE0 | At5g51550 | Hypothetical protein | 37.2 | 8.9 |
| | Q9LZQ4 | At3g62360 | Hypothetical protein | 133.8 | 5.99 |
| | Q9M1J5 | At3g57030 | Hypothetical protein | 40.9 | 7.71 |
| | Q9FMQ5* | At5g22460 | Hypothetical protein | 39.4 | 7.14 |
| | Q9ZW87 | At2g43105 | Hypothetical protein | 60.1 | 5.20 |
| | Q93Z16 | At4g21150 | Unknown protein | 74.9 | 7.16 |
| | Q9LNU5 | At1g20220 | Unknown protein | 58.7 | 9.65 |
| | Q9SSB7 | At1g80240 | Unknown protein | 40.4 | 8.53 |
| Others | P46689 | At1g75750 | Gibberellin-regulated protein 1 | 114 | 9.4 |
| | Q8L9P8 | At4g15800 | Rapid alkalinization factor (RALF) | 13.1 | 9.41 |
| | Q9ZV34 | At2g28790 | Thaumatin-like cytokinin binding protein | 27.8 | 7.92 |
| | Q8LBC7* | At3g27280 | Putative prohibitin | 30.6 | 6.93 |
| | Q9C568 | At5g22200 | NDR1/HIN1(harpin-induced)-like protein | 23.9 | 9.14 |
| | Q9LY84 | At5g14450 | Early nodule-specific protein-like | 44 | 8.72 |
| Likely contaminants | Q9MAU6 | At1g04980 | Disulfide isomerase-related protein, putative | 48.3 | 5.91 |
| | Q9LQG5 | At1g35620 | Protein disulfide isomerase | 53.3 | 5.16 |
| | O22263 | At2g47470 | Putative protein disulfide-isomerase | 39.8 | 5.80 |
| | Q8LDT6* | At5g66680 | Oligosaccharyltransferase-like protein | 48.8 | 5.9 |
| | Q00917 | At5g49190 | Sucrose-UDP glucosyltransferase | 92.3 | 5.79 |
| | Q9ZUA0 | At2g01720 | Putative ribophorin I | 52.2 | 8.29 |
| | P29402 | At5g61790 | Calnexin homolog 1 precursor | 60.7 | 4.81 |

 

**Table 1.** Continued

| Family[b)] | UniProt Accession number[c)] | Locus (AGI) | Description | MW (kDa) | p*I* |
|---|---|---|---|---|---|
| | P31166 | At1g27450 | Adenine phosphoribosyltransferase 1 | 26.6 | 8.65 |
| | Q9M9M5* | At3g03060 | Putative 26S proteosome regulatory subunit | 70.99 | 9.19 |
| | Q41951* | At3g16240 | Aquaporin tonoplast intrinsic protein 2.1 | 25.1 | 5.3 |
| | Q9FV53 | At1g15390 | Peptide deformylase, mitochondrial precursor | 29.1 | 8.82 |
| | Q9FHT4 | At5g37310 | Endosomal protein-like | 68.7 | 7.15 |
| | Q9LEV5 | At5g10840 | Putative endosomal protein | 73.8 | 8.15 |
| | Q9ZPS7 | At2g01970 | Putative endosomal protein | 68.6 | 6.65 |
| | P83291 | At5g20080 | NADH-cytochrome b5 reductase-like | 36.1 | 8.76 |
| | Q9ZNT1 | At5g17770 | NADH-cytochrome B5 reductase | 31.7 | 8.57 |
| | Q9ZS01 | At1g72730 | RNA helicase | 50.2 | 5.18 |

a) By definition all the proteins annotated as hypothetical proteins in this work should now be considered as expressed proteins.
b) The classification into functional classes was done based on UniProt annotation or/and on homology using Pfam.
c) Proteins found in both of the MudPIT experiments (protocol 1 and 2) are indicated by an asterisk.

26 genes encoding for α-expansin and five for β-expansin. In addition, there is a third group of related genes named expansin-like or expansin-related proteins comprising three and one member, respectively, in *Arabidopsis* (http://www.bio.psu.edu/expansins/arabidopsis.htm). In our work, two-expansin-like and three α-expansin proteins were identified.

Other well-known cell wall enzymes such as peroxidases were also represented in our data. Plant peroxidases are grouped in a superfamily, along with fungal and bacterial peroxidases, and contain three distantly related structural classes. Numerous peroxidase isozymes exist in plants and are mainly localised in the extracellular matrix and vacuole [35]. Class III contains all the plant peroxidases targeted to the secretory pathway and class I is composed of chloroplastic, mitochondrial, peroxisome and cytosolic ascorbate peroxidases [36]. In growing plant cell walls, peroxidases have been shown to promote the oxidative cross-linking of cell wall polymers [37]. Most of the peroxidases are basic proteins and are generally absent from 2-DE analysis; usually only acidic peroxidases are identified by this method [3]. Six peroxidases were identified in our list of wall proteins, all of them being basic proteins.

The primary cell wall typically contains glycosylated proteins such as hydroxyproline-rich glycoproteins (HRGP), often referred to as extensins. Recently, a novel family of HRGP, called LRR-extensin proteins (LRX) has been identified [38, 39]. Three members of this family AtLRX2, AtLRX3 and AtLRX4 (accession number O48809, Q9T0K5 and Q9LHF1, respectively) were found. The LRX family, which comprises 11 genes in *Arabidopsis*, encodes proteins characterised by a short N-terminal domain, a central region containing ten LRR domains, a cysteine-rich motif and a variable C-terminal extensin-like domain [38]. Whereas the N-terminal and LRR domains are well conserved amongst members of LRX family, the extensin domain was found to be highly variable. The extensin domain of AtLRX1 has been proposed to direct the protein to a particular region of the wall and/or to insolubilise it [39]. AtLRX3 and AtLRX4 are two paralogous genes that are members of the reproductive LRX subclass. They are expressed in all organs, although AtLRX4 expression was found to be higher in roots and young leaves than AtLRX3 [38]. Study of the related AtLRX1 gene suggests that this class of proteins could be involved in cell morphogenesis [39] and that LRX could function in the modification of localised cell wall domains during tissue differentiation. The functional analysis also revealed the presence of 17 potential contaminant proteins reducing the number of cell wall protein candidates to 89.

Certain classical proteins, such as arabinogalactan glycoproteins (AGP), glycine-rich (GRP), HRGP, or wall-associated kinases (WAK) are not present or are underrepresented in our data. Most of these proteins are heavily glycosylated proteins and therefore more soluble than other cell wall proteins. It is possible that they were washed away during the cell wall preparation because of the stringent washing conditions. Additionally, the glycosylation of polypeptides may reduce the chance for their identification by MS.

To summarise, the bioinformatic analysis identified 89 proteins that were either classical cell wall proteins or membrane proteins likely associated with the cell wall. These proteins varied in molecular weight from 8.9 to 133.8 kDa and had a spread of p*I* values (5.2 to 10.1). Seventy-six percent of the cell wall proteins were basic proteins concordant with the acidic environment of the cell wall (Fig. 1A and B). These values compared with the molecular weight from 4.2 to 334 kDa and p*I* between 4.14 and 12.24 for the total MudPIT list of 792 proteins (Fig. 1C and D).

 

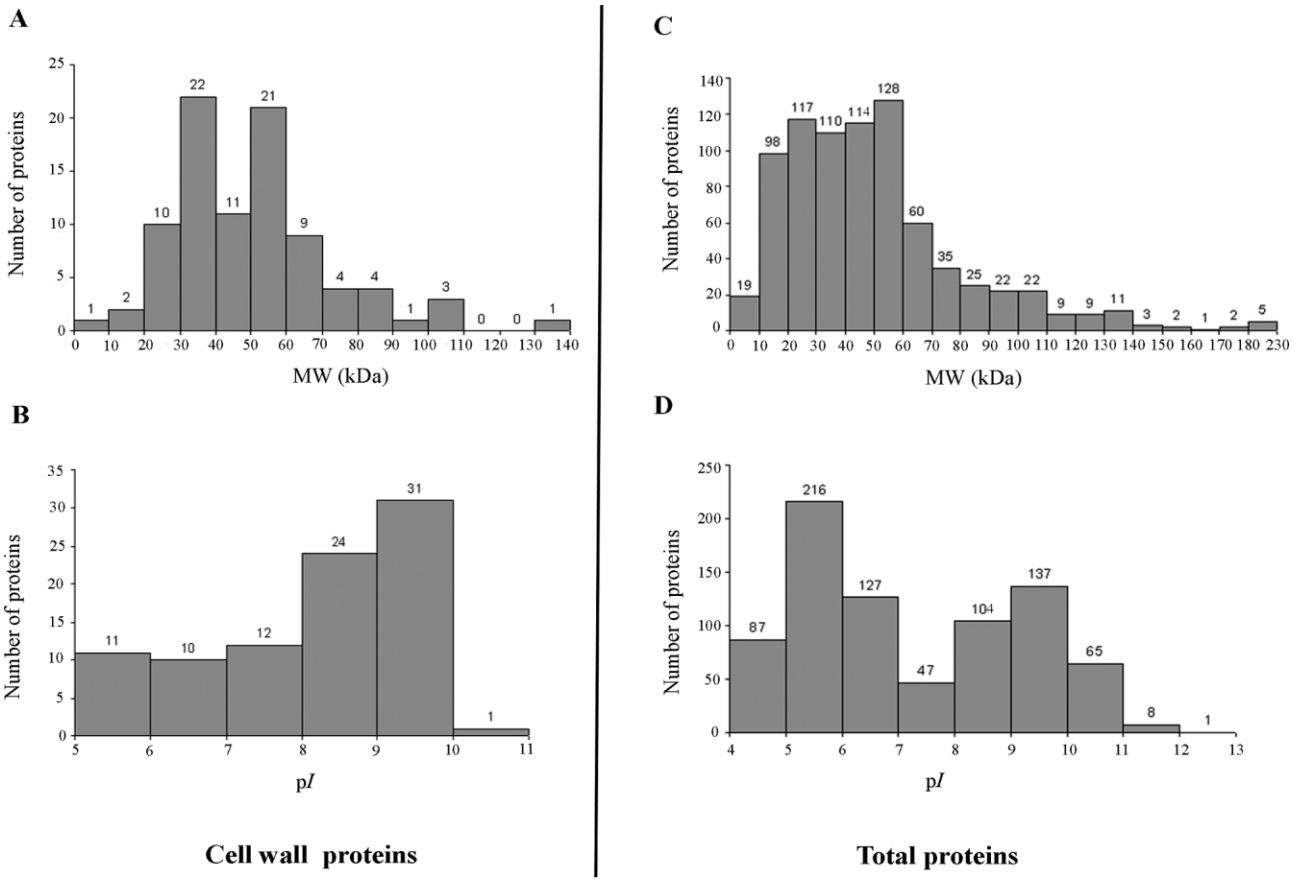


**Figure 1.** Distribution of the MW (A and C) and p*I* (B and D) of the selected cell wall proteins (A and B) compared with the total list of proteins (C and D) identified by MudPIT analysis of *Arabidopsis* cell wall preparations.

### 3.7 Comparison with previous cell wall and PM proteomic analyses

Our work adds to the previous proteomic analyses of the *Arabidopsis* cell wall [3–6]. In particular, we have focussed on the analysis of proteins not eluted by weak ionic solutions. Further, to maximise the potential for identifying all classes of proteins and proteins of low abundance we have used MudPIT. To assess the success of this different approach we compared our list of 89 cell wall proteins with the equivalent lists from the other published work (Table 2). Of the 89 proteins, 60 (67%) were unique to our study. As we would have predicted, the least overlap was with the studies of Charmont *et al.* [6] and Boudart *et al.* [4] both of which analysed the loosely-bound proteins eluted from the apoplastic environment of intact tissues. Recently, Marmagne *et al.* [40] reported a proteomic analysis of the plasma membrane from *Arabidopsis* suspension cells. About one hundred proteins were identified after SDS-PAGE and MS analyses. Of the 89 cell wall proteins we identified 3 (Q9M8T0, Q9ZU46, Q8LBC7) that were in common with those identified by Marmagne *et al.* [40]. Based upon our observations of the tight association of some membranous components with cell wall after purification [12], this was not unexpected. We contend that this close association justifies the inclusion of these proteins as *bona fide* wall proteins.

The current status of the proteomic analysis of *Arabidopsis* cell wall proteins is five independent publications that were based upon only a partial overlap in the technical resources and approaches followed. These differences are advantageous

**Table 2.** Overlap between published *Arabidopsis* cell wall proteomes

| Reference | This paper | [4] | [6] | [3] | [5][a)] |
|---|---|---|---|---|---|
| This paper | **89** | | | | |
| [4] | 11 | **87** | | | |
| [6] | 0 | 5 | **44** | | |
| [3] | 16 | 12 | 6 | **50** | |
| [5] | 14 | 7 | 1 | 11 | **41** |

a) From 'Chivasa *et al.* 2002' list of cell wall proteins, only the ones predicted to contain a signal peptide were taken into consideration for the comparison with other published cell wall proteomes.

 

in maximising the distinctiveness of the data sets obtained. To obtain a unified picture of the progress to date in this field we created a total list from all five proteomic studies of the *Arabidopsis* cell wall using the AGI gene notation, and eliminated redundant candidates to obtain the complete wall proteome list (Supplementary Data 6). For the work of Chivasa *et al*. [5], only genes for proteins carrying a signal peptide were included. The total list to date includes 268 unique genes, of which 60 (22%) were uniquely identified in our analysis of tightly-bound wall proteins. The limited level of redundancy between the five lists suggests that the published cell wall proteome has not yet reached saturation.

## 4 Concluding remarks

We have analysed the tightly-bound cell wall proteome for the primary cell wall from *Arabidopsis* using MudPIT. This sensitive and relatively unbiased approach identified a range of proteins with diverse physicochemical and biological properties. By using bioinformatics we identified secreted proteins that expanded the confirmed wall proteome by a further 60 unique proteins (22%). The low level of redundancy between all of the related studies suggests that the 268 unique proteins that comprise the *Arabidopsis* cell wall proteome is some way from saturation.

*We thank Keith Roberts for helpful comments during the progress of this work and for comments on the manuscript prior to submission. Proteomic analyses were carried out at the Joint IFR-JIC-UEA Proteomics Facility which was funded in part by: BBSRC JREI grant numbers JRE10832, JE412701, JE412631, HEFCE SRIF2 and by Syngenta and Unilever. EB was in receipt of a John Innes Foundation Postgraduate Studentship. The John Innes Centre is grant-aided from the United Kingdom Biotechnology and Biological Science Research Council (BBSRC).*

## 5 References

[1] Lee, S. J., Saravanan, R. S., Damasceno, C. M., Yamane, H. *et al.*, *Plant Physiol. Biochem.* 2004, *42*, 979–988.

[2] Cosgrove, D. J., *Annu. Rev. Cell Dev. Biol.* 1997, *13*, 171–201.

[3] Borderies, G., Jamet, E., Lafitte, C., Rossignol, M. *et al.*, *Electrophoresis* 2003, *24*, 3421–3432.

[4] Boudart, G., Jamet, E., Rossignol, M., Lafitte, C. *et al.*, *Proteomics* 2005, *5*, 212–221.

[5] Chivasa, S., Ndimba, B. K., Simon, W. J., Robertson, D. *et al.*, *Electrophoresis* 2002, *23*, 1754–1765.

[6] Charmont, S., Jamet, E., Pont-Lezica, R., Canut, H., *Phytochemistry* 2005, *66*, 453–461.

[7] Robertson, D., Mitchell, G. P., Gilroy, J. S., Gerrish, C. *et al.*, *J. Biol. Chem.* 1997, *272*, 15841–15848.

[8] Lin, D., Alpert, A. J., Yates, J. R., 3rd, *Am. Gen. Prot. Tech.* 2001, *1*, 38–46.

[9] Link, A. J., Eng, J., Schieltz, D. M., Carmack, E. *et al.*, *Nat. Biotechnol.* 1999, *17*, 676–682.

[10] Washburn, M. P., Wolters, D., Yates, J. R., 3rd, *Nat. Biotechnol.* 2001, *19*, 242–247.

[11] Koller, A., Washburn, M. P., Lange, B. M., Andon, N. L. *et al.*, *Proc. Natl. Acad. Sci. USA* 2002, *99*, 11969–11974.

[12] Bayer, E., Thomas, C. L., Maule, A. J., *Protoplasma* 2004, *223*, 93–102.

[13] Peck, S. C., Nuhse, T. S., Hess, D., Iglesias, A. *et al.*, *Plant Cell* 2001, *13*, 1467–1475.

[14] Neuhoff, V., Arold, N., Taube, D., Ehrhardt, W., *Electrophoresis* 1988, *9*, 255–262.

[15] Perkins, D. N., Pappin, D. J. C., Creasy, D. M., Cottrell, J. S., *Electrophoresis* 1999, *20*, 3551–3567.

[16] Bendtsen, J. D., Nielsen, H., von Heijne, G., Brunak, S., *J. Mol. Biol.* 2004, *340*, 783–795.

[17] Sonnhammer, E. L., von Heijne, G., Krogh, A., *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 1998, *6*, 175–182.

[18] Krogh, A., Larsson, B., von Heijne, G., Sonnhammer, E. L. L., *J. Mol. Biol.* 2001, *305*, 567–580.

[19] Eisenhaber, B., Wildpaner, M., Schultz, C. J., Borner, G. H. H., *et al.*, *Plant Physiol.* 2003, *133*, 1691–1701.

[20] Rice, P., Longden, J., Bleasby, A., *Trends Genet.* 2000, *16*, 276–277.

[21] Watson, B. S., Lei, Z., Dixon, R. A., Summer, L. W., *Phytochemistry* 2004, *65*, 1709–1720.

[22] Slabas, A. R., Ndimba, B., Simon, W. J., Chivasa, S., *Biochem. Soc. Trans.* 2004, *32*, 524–528.

[23] Verica, J. A., He, Z. H., *Plant Physiol.* 2002, *129*, 455–459.

[24] Möller, S., Croning, M. D. R., Apweiler, R., *Bioinformatics* 2001, *17*, 646–653.

[25] Ikezawa, H., *Biol. Pharm. Bull.* 2002, *25*, 409–417.

[26] Elortza, F., Nühse, T. S., Foster, L. J., Stensballe, A. *et al.*, *Mol. Cell. Proteomics* 2003, *2*, 1261–1270.

[27] Borner, G. H., Lilley, K. S., Stevens, T. J., Dupree, P., *Plant Physiol.* 2003, *132*, 568–577.

[28] Borner, G. H., Sherrier, D. J., Stevens, T. J., Arkin, I. T., Dupree, P., *Plant Physiol.* 2002, *129*, 486–499.

[29] Rose, J. K. C., Saladié, M., Catalá, C., *Curr. Opin. Plant Biol.* 2004, *7*, 296–301.

[30] Henrissat, B., Coutinho, P. M., Davies, G. J., *Plant Mol. Biol.* 2001, *47*, 55–72.

[31] Van Den Bulcke, M., *Proc. Natl. Acad. Sci. U.S.A.* 1989, *86*, 2673–2677.

[32] Cosgrove, D. J., *Plant Physiol. Biochem.* 2000, *38*, 109–124.

[33] Campbell, P., Braam, J., *Trend Plant Sci.* 1999, *4*, 361–366.

[34] Cosgrove, D. J., *Nature* 2000, *407*, 321–326.

[35] Mäder, M., in Penel, C., Gaspar, T., Greppin, H., (Eds), *Compartmentation of Peroxidase Isoenzymes in Plant Cells*, In: *Plant Peroxidases* 1980–1990, University of Geneva, Switzerland 1992, pp. 37–46.

[36] Welinder, K. G., *Curr. Opin. Struc. Biol.* 1992, *2*, 388–393.

[37] Fry, S. C., *Annu. Rev. Plant Physiol.* 1986, *37*, 165–186.

[38] Baumberger, N., Doesseger, B., Guyot, R., Diet, A. *et al.*, *Plant Physiol.* 2003, *131*, 1313–1326.

[39] Baumberger, N., Ringli, C., Keller, B., *Genes Dev.* 2001, *15*, 1128–1139.

[40] Marmagne, A., Rouet, M. A., Ferro, M., Rolland, N. *et al.*, *Mol. Cell. Proteomics* 2004 *3*, 675–691.

 