Plant cell wall proteomics: mass spectrometry data, a trove for research on protein structure/function relationships.

Cécile Albenne, Hervé Canut, Georges Boudart, Yu Zhang, Hélène San Clemente, Rafael Pont-Lezica, Elisabeth Jamet

Molecular Plant, Special Issue: Cell Wall, 2009, 2: 977-989.

Plant cell wall proteomics: mass spectrometry data, a trove for research on protein structure/function relationships

Cécile Albenne, Hervé Canut, Georges Boudart, Yu Zhang, Hélène San Clemente,
Rafael Pont-Lezica, Elisabeth Jamet[1]

Surfaces Cellulaires et Signalisation chez les Végétaux, UMR 5546 CNRS - UPS - Université de Toulouse, Pôle de Biotechnologie Végétale, 24 chemin de Borde-Rouge, BP 42617 Auzeville, 31326 Castanet-Tolosan, France

[1] To whom correspondence should be addressed. E-mail jamet@scsv.ups-tlse.fr, fax +33 562 193 502, tel. +33 562 193 530.

ABSTRACT

Proteomics allows the large scale study of protein expression either in whole organisms or in purified organelles. In particular, mass spectrometry (MS) analysis of gel-separated proteins produces data not only for protein identification, but for protein structure, location, and processing as well. An in-depth analysis was performed on MS data from etiolated hypocotyl cell wall proteomics of *Arabidopsis thaliana*. These analyses show that highly homologous members of multigene families can be differentiated. Two lectins presenting 93% amino acid identity were identified using peptide mass fingerprinting. Although the identification of structural proteins such as extensins or hydroxyproline/proline-rich proteins (H/PRPs) is arduous, different types of MS spectra were exploited to identify and characterize an H/PRP. Maturation events in a couple of cell wall proteins (CWPs) were analyzed using site mapping. *N*-glycosylation of CWPs as well as the hydroxylation or oxidation of amino acids were also explored, adding information to improve our understanding of CWP structure/function relationships. A bioinformatic tool was developed to locate by means of MS the *N*-terminus of mature secreted proteins and *N*-glycosylation.

Key words: cell wall protein; MALDI-TOF; mass spectrometry; post-translational modification; protein structure; proteomics

INTRODUCTION

The plant cell wall is an essential cell compartment that has structural as well as functional roles. The envelope of the plant cell is not only responsible for its shape; it is also a dynamic structure in direct contact with the environment. Mostly composed by polysaccharides (up to 95 % in mass) the cell wall also contains proteins, and in some cell types, lignins (Carpita and Gibeaut, 1993). The composition and organization of polysaccharides have been widely described. The global architecture of this network is now well studied, and several models are available (Willats et al., 2001; Fry, 2004; Somerville et al., 2004). Among cell wall proteins (CWPs), structural proteins such as extensins and hydroxyproline/proline-rich proteins (H/PRPs) are known to reinforce the polysaccharide networks (Showalter, 1993; Kieliszewski and Lamport, 1994). From a functional point of view, the cell wall is essential for growth and differentiation, signaling, and response to biotic and abiotic stresses (Roberts, 1990,2001; Ellis et al., 2002; Vogel et al., 2004). Since the true actors of cell wall dynamics are proteins, all CWPs other than structural proteins are of interest. Therefore, to better understand the cell wall complexity, the challenge is to go further into the identification of the CWPs and their structure/function relationships. In this context, the last few years saw the emergence of cell wall proteomics aimed at profiling CWPs at a given time in specific environmental conditions. Beyond some studies performed on *Medicago sativa* (Watson et al., 2004), *Zea mays* (Alvarez et al., 2006; Zhu et al., 2006), *Cicer arietinum* (Bhushan et al., 2006), and *Oryza sativa* (Chen et al., 2008; Jung et al., 2008; Cho et al., 2009) most data on plant cell wall proteomics concern the model plant *Arabidopsis thaliana*. The knowledge on CWPs has recently been reviewed and updated (Jamet E. et al., 2006; Jamet E et al., 2008). As an active contributor to the description of *A. thaliana* CWPs (Borderies et al., 2003; Boudart et al., 2005; Charmont et al., 2005; Feiz et al., 2006; Minic et al., 2007; Irshad et al., 2008), our group has built a database (*WallProtDB*) containing all *A. thaliana* and *O. sativa* CWPs identified in different organs. So far, this database contains nearly 800 CWPs and is freely accessible to the public (http://www.polebio.scsv.ups-tlse.fr/WallProtDB/). CWPs in the database were classified into nine functional classes according to bioinformatic predictions, thus revealing a great diversity of protein functions inside the cell wall. We now have a good picture of CWPs present in the plant cell wall, but the exact functions of most of these proteins are still unknown. Several descriptions of CWP mutants are proof of the importance of proteins in cell walls both during plant development and in response to pathogens (Berger and Altmann, 2000; Baumberger et al., 2003; Vogel et al., 2004; Roudier et al., 2005; Kurdyukov et al., 2006). However, many topics related to CWPs still need to be addressed to achieve a better understanding of cell wall dynamics. Among these topics are: the role of different members of multigenic families, the identification of substrates or partners, the post-translational modifications (PTMs), and the spatial and temporal regulation.

Several strategies can be used in cell wall proteomic studies (Jamet E et al., 2008). A simplified flowchart for plant cell wall proteomics is presented in Figure 1. Defining a specific strategy to answer a biological question requires two essential decisions involving: (i) the choice of the appropriate material; and (ii) a method to prepare a protein extract containing as few intracellular proteins as possible (Feiz et al., 2006). Since CWPs are mostly basic glycoproteins, a good protein separation was obtained simply by mono-dimensional electrophoresis (1D-E) (Boudart et al., 2005). However, to improve CWP fractionation, cationic exchange chromatography (CEC) performed in acidic pH was shown to be a method of choice, because it mimics the physico-chemical conditions in cell walls (Irshad et al., 2008). In a second step, chromatography fractions were separated by 1D-E, thus offering a more suitable bi-dimensional (2D) separation than 2D-E. The staining of gels can be done in

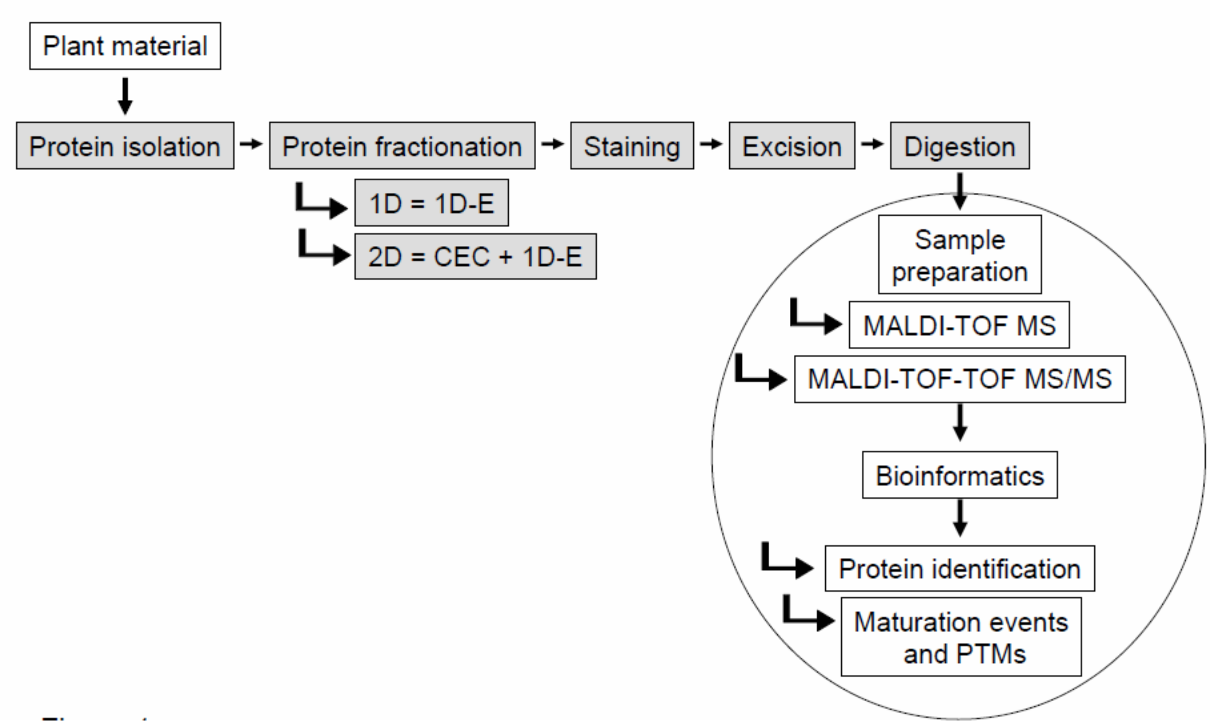

Figure 1. A simplified plant cell wall proteomic flowchart.

Three main steps are required to obtain information on proteins using MS from plant material. The first step is the collection of plant material. The second step is protein extraction and separation. Efficient fractionation of CWPs can be achieved in different ways: 1D-separation (monodimensional) only using 1D-E (mono-dimensional electrophoresis); or 2D-separation (bi-dimensional) using cationic exchange chromatography (CEC) followed by 1D-E of eluted fractions. In the third step, after staining, excision, protease digestion, and sample preparation, MS analysis is performed, followed by bioinformatic data processing. This leads to a flowchart that allows protein identification and gives information on maturation events and PTMs.

different ways, revealing different types of proteins with different sensitivities (Chevalier et al., 2004; Jamet E et al., 2008). Stained bands of interest were excised from gels and subsequently submitted to protease digestion prior to analysis by MS. Two approaches are widely used for protein identification: (i) peptide mass fingerprinting (PMF) (MS spectrum) and (ii) peptide sequencing by tandem mass spectrometry (MS/MS spectrum) (Aebersold and Mann, 2003). The main advantage of the flowchart in Figure 1 is that it provides samples containing whole proteins ready for MS analysis. Analysis of complex samples was also performed, but is more difficult to interprete because the peptides of all the proteins are mixed (Bayer et al., 2006). Reconstruction of whole proteins requires more powerful software, and the structural information for a given protein is more difficult to obtain. Each step described in Figure 1 can be optimized to yield better results. In addition, MS technologies have evolved very quickly these last years, and they offer a large choice of sources of ionization and of analyzers (Aebersold and Mann, 2003; Han et al., 2008). However, to identify and characterize proteins, MALDI-TOF MS and MALDI-TOF-TOF MS/MS remain powerful, easy to use methods. They are the only two techniques to be considered in this paper. Finally, the bioinformatic analysis is critical because it allows not only protein identification, but also prediction of protein sub-cellular location and functional domains (San Clemente et al., 2009).

Although proteomics has been shown to be a very valuable way to identify CWPs, mass spectrometry (MS) data can yield more information on protein structure, location of PTMs and processing, thus tackling biochemical protein complexity and revealing the hidden part of the iceberg. This article seeks to reveal the potential capacity of MS to improve our knowledge of CWPs. Relevant examples were chosen from a previous cell wall proteomic study in etiolated *A. thaliana* hypocotyls, where 137 CWPs were identified (Irshad et al., 2008). A detailed analysis of the data allowed the discrimination of close members of multigenic families, highlight maturation events, locate PTMs, or access proteins which are recalcitrant under standard conditions.

RESULTS AND DISCUSSION

Discrimination among close members of multigenic families

Most plant genes including those encoding CWPs, belong to multigene families. These can be very large, and they can include closely related members. For example, in *A. thaliana* there are 73 predicted peroxidase genes (Oliva et al., 2009), 66 pectin methylesterase (PME) genes (http://cellwall.genomics.purdue.edu/families/4-5-1.html), 33 xyloglucan endotransglucosylase/hydrolase genes (Rose et al., 2002) and 66 polygalacturonase genes (http://cellwall.genomics.purdue.edu/families/4-3-3.html). It is usually difficult to study the regulation of close members of these gene families at the protein level. Classical approaches using immunology are not efficient because of high amino acid similarity, but PMF using MALDI-TOF MS provides a powerful alternative. Indeed, subtle mass differences between peptides can be measured, the accuracy of mass determination being as small as 20 ppm (parts per million). As an example, the identification in the same sample of two closely related lectin homologues is reported in Figure 2. At1g78850 and At1g78860 share 93% amino acid identity. On MS spectra, seven peptides common to both proteins were found (boxed peptides) whereas four and three peptides were respectively specific to At1g78850 (underlined in green) and At1g78860 (underlined in blue) (Supplementary Figure 1). Note that in one case, the two versions of a peptide could be easily distinguished (ILENGNMVIYDS**S/N**GK), that is, a difference of 11.02 Da in mass (1655.79 Da and 1666.81 Da respectively). In our study, all CWPs could be identified unambiguously, but some intracellular contaminants of our cell wall preparation could not be discriminated because they belong to multigene families comprising closely related members. Either there was not enough protein coverage to find specific peptides, or the amino acid sequences of the proteins were too close, *e.g.* malate dehydrogenase (88% shared identity between At1g53240 and At3g15020), RUBISCO small subunit (96% or 97% shared identity among At5g38410, At5g38420, and At5g38430), actin (100% shared identity between At2g37620 and At3g53750). The last case was that of proteins with repetitive domains such as polyubiquitins and ubiquitin extension proteins (*e.g.* At3g52590, At2g36170, At2g35635, and At1g31340).

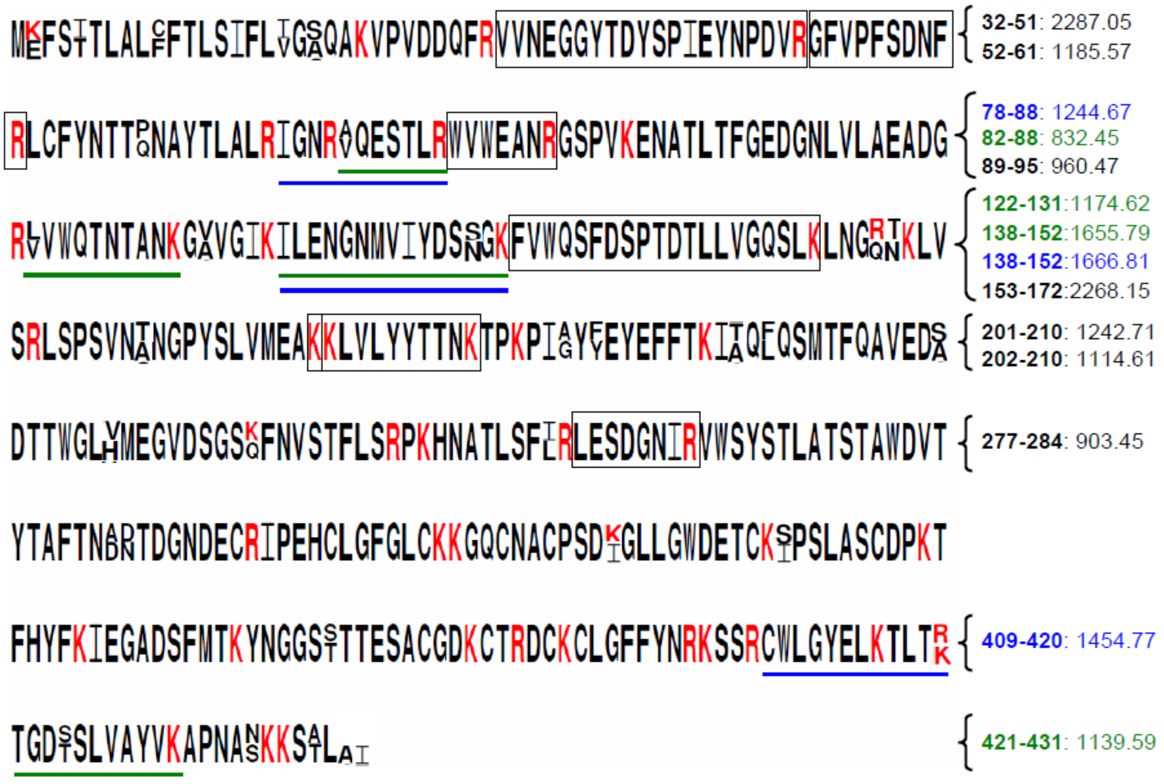

Figure 2. Identification of members of multigene families by peptide mass mapping.

The amino acid sequences of two *A. thaliana* lectin homologues *At1g78850* (NP_178006) and *At1g78860* (NP_178007) were superimposed using the WebLogo 3 software (http://weblogo.threeplusone.com/). Common amino acids are identified with a single capital letter and differences in amino acid sequences appear in small superimposed letters (upper line for NP_178006, lower line for NP_178007). Black frames express common peptides found by protein PMF. Peptides discriminating between NP_178006 and NP_178007 are respectively underlined in green and blue. The positions and the monoisotopic masses of the peptides are written on the right in black, green, and blue for peptides common to both proteins, specific to NP_178006, and specific to NP_178007, respectively.

A site mapping approach to post-translational maturation events

To become biologically active, some enzymes need to lose a pro-sequence, and are submitted to post-translational maturation events. PMF can be used as a tool to show such events by mapping the identified peptides on the sequence of the proteins. To be reliable, it must be carried out on several independent samples. Two examples of possible CWP maturation events are shown in Figure 3.

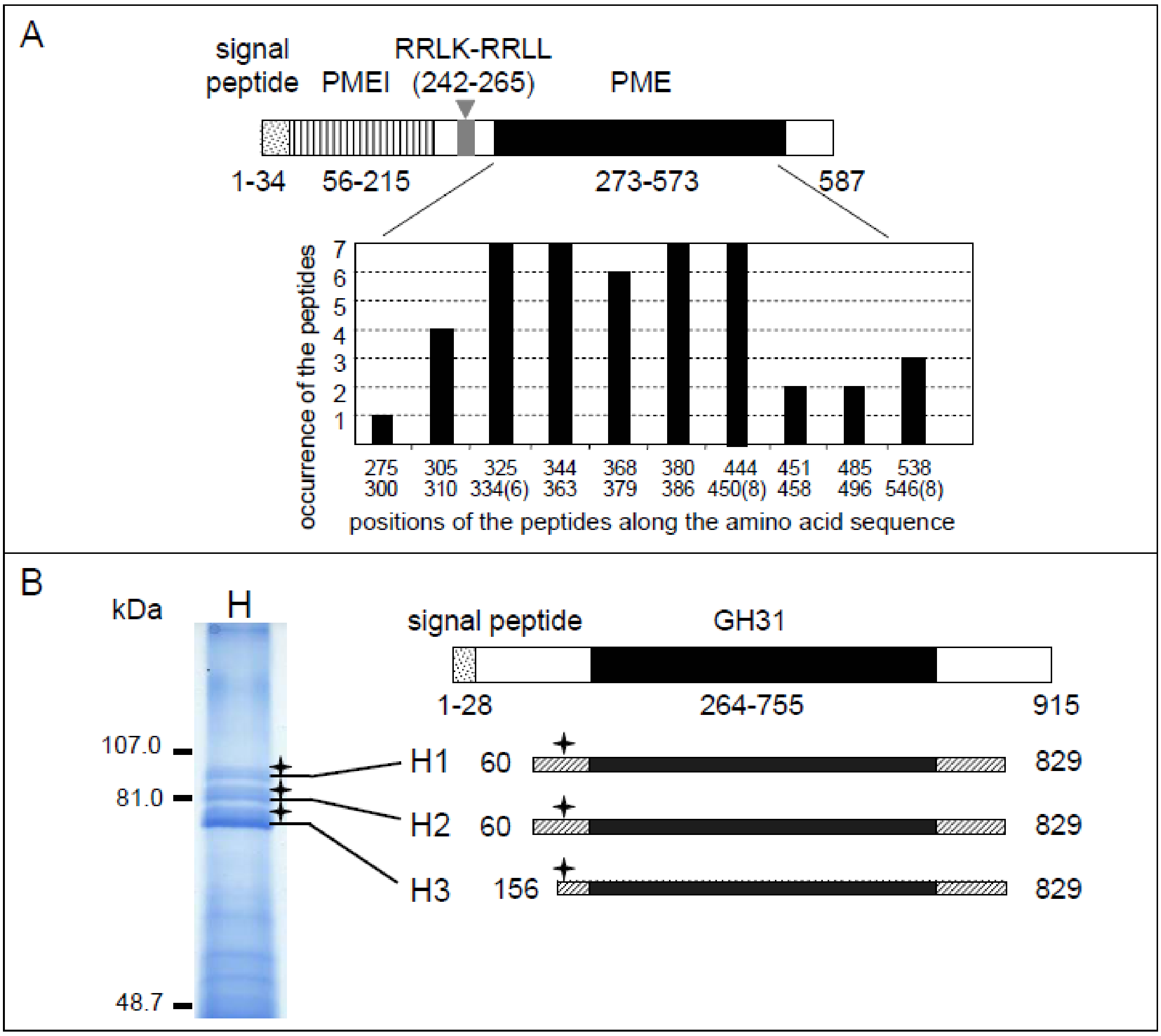

Figure 3. Looking for post-translational maturation events: *At1g53830* (NP_175786) encoding a pectin methylesterase (**A**) and *At1g68560* (NP_177023) encoding an α-xylosidase (**B**) as test cases.

In **A**, At1g53830 is represented as a scheme showing four predicted domains: a signal peptide (between amino acids 1 and 34, dotted box); a pectin methylesterase inhibitor domain (PMEI, between amino acids 56 and 215, striped box); a pectin methylesterase domain (PME, between amino acids 273 and 573, black box). RRLK and RRLL located between amino acids 242 and 265 (grey box) are the two conserved basic tetrad motifs predicted to be essential for the cleavage of the PMEI domain. The occurrence of peptides within the PME domain in seven independent PMF experiments is indicated. In **B**, At1g68560 is represented as a scheme showing two predicted domains: a signal peptide (between amino acids 1 and 28, dotted box); and a glycosyl hydrolase family 31 (GH31) domain (between amino acids 264 and 765). Six samples were taken from the H fraction separated by 1D-E after CEC described in Irshad *et al.* (2008). The limits of PMF sequence coverage of the protein in samples H1, H2 and H3 are indicated. Stars indicate the position of an *N*-glycosylated peptide only found in the upper parts of bands H1, H2 and H3. In **A** and **B**, white boxes on the schemes of the proteins correspond to regions where no specific domain was predicted.

In our proteomic study on etiolated *A. thaliana* hypocotyls, six different putative pectin methylesterases (PMEs) were identified (Irshad et al., 2008). Five of them are type-I PMEs with both a predicted *N*-terminal pectinesterase inhibitor domain (PMEI, PF04043), and a *C*-terminal pectinesterase domain (PME, PF01095) (Micheli, 2001). The cleavage of the pro-sequence was shown to depend on the presence of one or two conserved basic tetrad motifs (Wolf et al., 2009). In none of the cases did we find peptides in the PMEI domains. In seven independent identifications of At1g53830 (Figure 3A), all the peptides were mapped to the predicted PME domain, *i.e.* between amino acids 275 and 548 thus allowing a coverage of 47% of this region (Supplementary Figure 2). In addition, the apparent molecular masses of these PMEs after 1D-E in denaturing conditions was around 35 kDa, consistent with a maturation event of the protein. For a long time, it was thought that there was no PMEI domain in cell wall PMEs. It was assumed that PMEI domains of PMEs could inhibit their enzyme activity thus preventing the premature de-esterification of pectins before their secretion (Micheli, 2001). However, it was recently shown that maturation of PMEs inside the secretory pathway is required for their secretion (Bosch et al., 2005; Wolf et al., 2009). Our results are consistent with the single presence of mature PMEs in cell walls. They strongly support the assumption that maturation occurs before protein secretion. However, this must be confirmed by Edman degradation sequencing to provide the actual *N*-terminus of the processed protein.

The second example of protein maturation is provided by α-xylosidase At1g68560. In our samples, α-xylosidase was repeatedly identified as the only component present in three wide Coomassie blue stained bands obtained after 1D-E in denaturing conditions (H1, H2, H3 in Figure 3B) (Irshad et al., 2008). Apparent molecular masses were between 75 and 100 kDa. Cartography of identified peptides on 18 independent fractions indicates that proteins in H3 lack *N*-terminal peptides (Supplementary Figure 3). Differences in molecular masses between H1, H2, and H3 are thus probably due either to the *N*-terminal processing of the protein, or to *N*-glycosylations. In fact, there are eight predicted *N*-glycosylation sites in the protein. We found that at least one of these sites is occupied. Stars in Figure 3B point to *N*-glycosylated forms of the protein on the SNHETLF**NTTS**SLVFK peptide located between amino acids 156 and 173. In addition, At1g68560 was identified in a previous proteomic study performed on a stem protein fraction retained on Concanavalin A (Minic et al., 2007). We suspect that different states of *N*-glycosylation may occur on the other predicted sites. Furthermore, our MS data were compared to results of the *N*-terminal sequencing of a cabbage α-xylosidase, which is also a member of *Brassicaceae*. This enzyme shows 75% identity with At1g68560 and an apparent molecular mass of 85 kDa after 1D-E in denaturing conditions (Sampedro et al., 2001). Edman sequencing showed this *N*-terminal sequence was ISGSELTFSYTTDPFSFAVKRRL (J Sampedro, personal communication). It shares 87% identity with an At1g68560 peptide located between amino acids 123 and 146. The removal of the peptide upstream of this sequence would cause a mass shift of 10.9 kDa, which fits in with the difference observed between molecular masses of H2 and H3. Altogether, it suggests that At1g68560 is processed at its *N*-terminus in cell walls. The processing of cell wall enzymes that act on polysaccharides was already shown. An α-L-arabinofuranosidase (ARAI) and a β-D-xylosidase (XYL) from barley are post-translationally processed at their *C*-terminus (Lee et al., 2003). Such *C*-terminal processing was also suggested for the β-D-xylosidases AtXYL1 and AtXYL4 (Minic et al., 2004). In this case, not all the *C*-termini of the proteins were covered by MALDI-TOF MS PMF. An endo-β-1,4-xylanase (XYN-1) of barley aleurone was shown to be processed both at its *N*- and *C*-termini (Caspers et al., 2001). In addition, cell wall proteomic studies have shown that proteases represent about 11.6% of identified CWPs, none of which have known substrates (Jamet E et al., 2008). In conclusion, our data support the hypotheses of an *N*-terminal processing of the At1g68560 α-xylosidase *in muro* and of the presence of *N*-glycans.

In order to facilitate MS data analysis and easily identify post-translational maturation or modification events, we developed a bioinformatic tool called *ProTerNyc*, which is freely accessible (http://www.polebio.scsv.ups-tlse.fr/ProTerNyc/). This software predicts the location of the *N*-terminus in mature secreted protein and the *N*-glycosylated peptides using the data obtained by MS. The *N*-terminus is predicted by TargetP (Emanuelsson et al., 2007). With this tool it is possible to overcome the limitations of searches in databases that do not take into account the signal peptide cut or peptide glycosylation. After running *ProTerNyc* on a given protein sequence and on a mass list generated by MS experiments, it can be seen that some observed peaks can be attributed to putative *N*-terminal peptide and *N*-glycopeptides.

Location of post translational modifications

Aside from all the PMF data used for the cartography of a protein sequence, it is sometimes very instructive to go further into the MS spectra. When exploring the MS data for etiolated hypocotyl CWP fractions (Irshad et al., 2008), we observed some significant peaks not always used for protein identification. A detailed analysis of these data revealed that they are consistent with PTMs such as *N*-glycosylation or amino acid hydroxylation or oxidation.

N-glycosylations

Several peroxidases were found (Irshad et al., 2008). Figure 4A shows the MS spectrum allowing the identification of one of these peroxidases, encoded by *At3g32980*. Among the major peaks displayed in this spectrum, five were found to fit with non-modified peptides of At3g32980, leading to its identification. Other peaks in this spectrum were shown to belong to other proteins also identified in this sample, except for four peaks of high m/z (2661.16, 2864.25, 3685.54, and 3888.65). Since At3g32980 carries four putative *N*-glycosylation sites, we looked for possible glycopeptides. Using the m/z of the four peaks listed above (red arrows in Figure 4A) and those of canonical *N*-glycans observed in plants (Faye et al., 2005), we calculated that each of these four peaks may correspond to two modified peptides, P1 and P2. Both peptides possess one *N*-glycosylation consensus sequence and are supposed to carry two different complex motifs of *N*-glycosylation. Interestingly, in their unmodified form P1 and P2 were absent from the spectrum suggesting that in our experimental conditions these glycosylation sites are always occupied. To confirm that these four peaks correspond to glycopeptides, we analyzed a total protein extract after hydrogen fluoride (HF) deglycosylation. MS analysis of the corresponding sample showed that all four peaks were absent (Figure 4B). This suggests that glycosylations were removed during the HF treatment, confirming the nature of these modifications.

It has been reported that *N*-glycans are asymmetrically distributed on soybean seed coat peroxidase, whereas they are more uniformly distributed over horseradish peroxidase (Gray, 1998; Gray and Montgomery, 2006). In addition, a study performed on cationic peanut peroxidase showed that *N*-glycans are essential to attain full catalytic ability and thermostability, and to influence protein folding (Lige et al., 2001). Other examples demonstrate the importance of protein *N*-glycosylation in plants. The presence of *N*-glycans on the jack bean α-mannosidase was shown to influence the enzyme activity as well as its oligomerization (Kimura et al., 1999). It was also proposed that *N*-glycosylation motifs on a plant cell wall polygalacturonase inhibitor protein (PGIP) may favor interaction with its fungal polygalacturonase target during host-pathogen interaction (Lim et al., 2009). Finally, several studies on defective glycosylation mutants have shown that the *N*-glycosylation pathway regulates protein quality control, salt tolerance, cell wall deposition, development, and plant-pathogen interactions (Pattison and Amtmann, 2009).

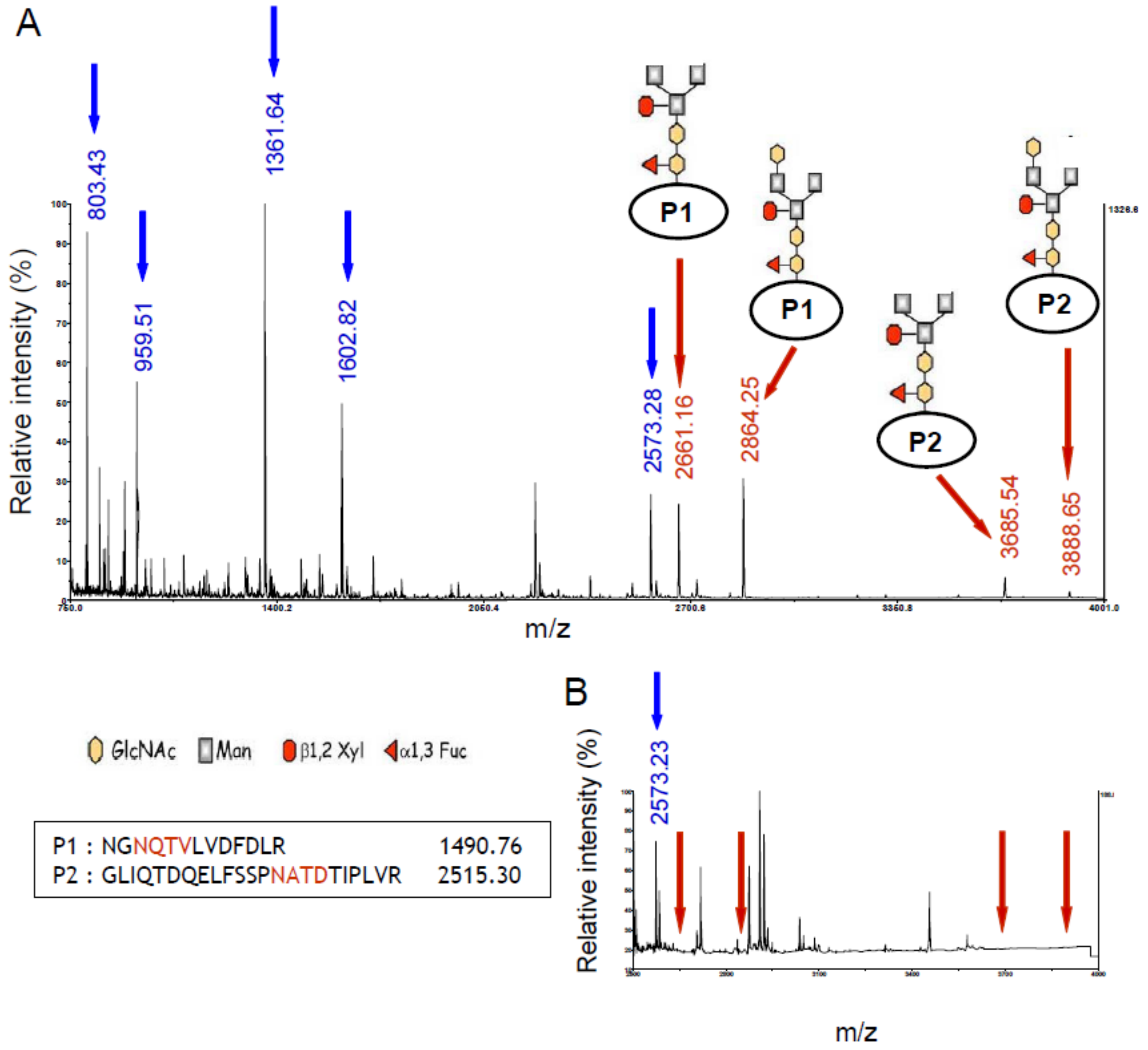

Figure 4. Identification of *N*-glycopeptides on At3g32980 (NP_850652) peroxidase by MALDI-TOF MS analysis.

The protein was identified in the A CWP fraction obtained by CEC (Irshad et al., 2008) (**A**) and in a total CWP extract deglycosylated by hydrogen fluoride (**B**). Non modified peptides used for identification are shown with blue arrows and *N*-glycopeptides with red arrows. Two peptides (P1 and P2) carrying *N*-glycosylation consensus sequences (in red in the frame) were shown to carry *N*-glycosylations. The mass of these *N*-glycopeptides corresponds to the addition of the mass of the peptide and that of the consensus *N*-glycans represented schematically as described (Faye et al., 2005). A zoomed spectrum (m/z from 2500 to 4000) of the deglycosylated sample shows that glycopeptides are absent.

Our results show that MS can be used to predict glycosylation and to gain new insight into the structural and functional characterization of proteins. However, until now, it was difficult to recognize *N*-glycopeptides using a systematic proteomic approach. Software such as ProteinProspector (http://prospector.ucsf.edu/) and MASCOT (http://www.matrixscience.com/home.html) provide a series of simple PTM options to search for peptides with modified masses, including modified amino acids. In-depth PTM analysis, such as glycosylations, requires the use of specific tools since this information is not yet referenced in databases. For a given protein, the *ProTerNyc* tool (http://www.polebio.scsv.ups-tlse.fr/ProTerNyc/) described above can be used to calculate all putative *N*-glycopeptides and to identify them in a mass list obtained from MS experiments. It should be noted that mass tolerance for peptides with high m/z may need to be increased since accuracy of mass determination decreases as mass increases. This software will undoubtedly be a precious tool to search for glycopeptides.

Amino acid hydroxylation or oxidation

Many CWPs contain hydroxylated Pro (Hyp), which can be *O*-glycosylated (Kieliszewski and Lamport, 1994). The At3g16850 polygalacturonase was identified by PMF using twelve peptides among CWPs from etiolated hypocotyls (Irshad et al., 2008). Among these peptides, one (P peptide) was predicted to exist under several forms. Indeed, the mass spectrum displays an isotopic massif with a monoisotopic peak at m/z 2522.23, associated with two isotopic massifs at a distance of 16 Da and 32 Da respectively (Figure 5). Such modifications may respectively correspond to the oxidation or hydroxylation of one or two amino acids. The P peptide has a theoretical monoisotopic mass of 2522.27 Da (Figure 5). We performed MALDI-TOF-TOF MS/MS experiments on each of these peptides in order to confirm that these three isotopic massifs correspond to the same peptide, and locate oxidations or hydroxylations of amino acids. A clear fragmentation pattern was obtained from the P peptide, confirming the proposed sequence (Supplementary Figure 4). Both MS/MS spectra of the modified forms (m/z 2583.23 and 2554.21) present the two most abundant fragments previously observed for the P peptide (y10 and y15, Supplementary Figure 5). However, one of these fragments (y15) differed by 16 Da or 32 Da from the corresponding peaks found on the P peptide fragmentation pattern. These results confirmed the presence of oxidations or hydroxylations and allowed their localization in a region of 5 possible amino acid residues (PSHWD, in bold in Figure 5). The oxidized or hydroxylated amino acids in this sequence probably are the Pro, His, Trp and Asp, which respectively give rise to Hyp, oxohistidine, hydroxytryptophan and hydroxyaspartate. Hyp is widespread in the plant kingdom. Pro hydroxylation is a PTM performed by prolyl 4-hydroxylases in the endoplasmic reticulum (Lamport, 1965). A code based on primary sequences was proposed for Pro hydroxylation and consecutive *O*-glycosylation in Hyp-rich glycoproteins (HRGPs) and in Hyp/Pro-rich proteins (H/PRPs) (Shpak et al., 1999; Kieliszewski, 2001). In the case of At3g16850, only one Pro may be hydroxylated in the P peptide within the Asp-Pro-Ser context which has never been described in plant proteins. In addition, the presence of Hyp in a CWP different from the well-known HRGPs and H/PRPs is noteworthy. A few cases have been reported in the literature, such as that of the dodeca-CLE peptides involved in cell differentiation (Ito et al., 2006; Kondo et al., 2006). Hyp was found in the Val-Pro-Ser and Gly-Pro-Asn sequences. So far, Pro hydroxylation in the CLE peptides seems to have no effect on their biological activity. Unlike Hyp, the presence of oxohistidine, hydroxytryptophan, and hydroxyaspartate in a plant protein has never been reported. Several examples of metal-catalyzed oxidation of His to oxohistidines were described for human proteins (Schöneich, 2000; Hovorka et al., 2002). Hydroxytryptophan is described as a precursor of the neurotransmitter serotonin and as an intermediary in tryptophan metabolism (Birdsall, 1998). Hydroxyaspartate residues were found in epidermal growth factor domains of several mammalian proteins. It was proposed that this modification regulates negatively the fucosylation of these regions (Castellino et al., 2008). In plants, oxidative modifications are known to occur in presence of reactive oxygen species and metals, in particular in cell walls (Møller et al., 2007). In the case of At3g16850, suspected oxidized or hydroxylated amino acids do not belong to the active site according to bioinformatic prediction. We can note that in the closest plant homologs of At3g16850 the pentapeptide DPSHW is conserved, suggesting its importance for protein function. However, since no structural data are available and no catalytic residues have been identified for any plant polygalacturonase, it is still difficult to predict a role for these amino acids.

P peptide : GATILASPD**PSHWD**VVSPLPSYGR m/z = 2522.27

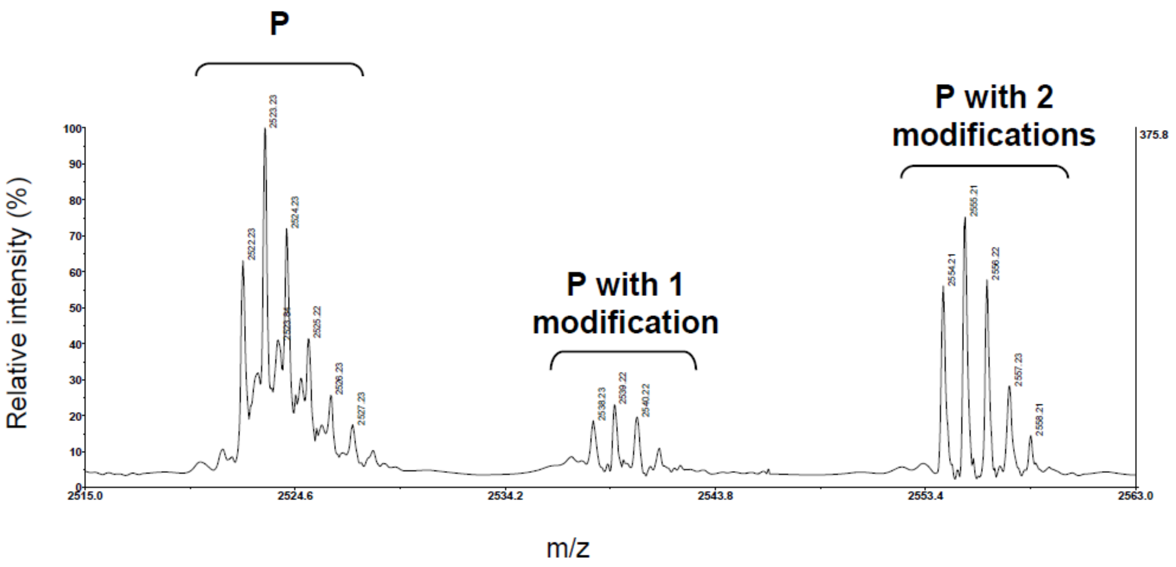

Figure 5. Location of amino acid oxidation or hydroxylation on the At3g16850 (NP_188308) polygalacturonase by MALDI-TOF MS analysis.

The protein was identified in the Q CWP fraction obtained by CEC (Irshad et al., 2008). PMF identification was achieved with 12 peptides among which the P peptide (sequence and mass indicated in frame). The range of the spectrum zoom shown (m/z between 2515 and 2563) corresponds to the range where P is observed, with no modification or with one or two modifications (mass shifts of 16 Da and 32 Da respectively). The two amino acid modifications are assumed to occur within the sequence indicated in bold.

Identification and characterization of recalcitrant structural CWPs

In all the plant cell wall proteomes described so far, structural cell wall glycoproteins such as extensins and H/PRPs represent only 1.7 % of the presently identified CWPs and remain under-represented (Jamet E et al., 2008). Many reasons may explain this situation: (i) the extraction of structural proteins is difficult since they can be strongly bound to the cell wall; (ii) the separation by electrophoresis in denaturing conditions can be disturbed by the presence of glycosylations; (iii) digestion into peptides suitable for standard MS experiments can be limited because of the low frequency of protease cut sites into repetitive sequences, rendering protein identification by PMF difficult. Alternative strategies must be used to identify and characterize these proteins, such as peptide sequencing by tandem MS and MS in linear mode.

The At5g14920 H/PRP, also named AtGASA14 (Roxrud et al., 2007), was identified in etiolated hypocotyls and a smear was observed after 1D-E, with an apparent molecular mass between 35 and 40 kDa (between K14 and K11 bands in Figure 6A) (Irshad et al., 2008). Aside from the Pro-rich region covering most of the protein, At5g14920 also contains a cleavable *N*-terminal signal peptide and a *C*-terminal sequence which shows homology with the GASA peptides previously described in plants (Figure 6B). Analysis of the primary sequence of the Pro-rich region indicates that it contains only two tryptic cut sites since this protease cannot cleave Lys-Pro linkages. Altogether, in the entire sequence, only 12 Lys/Arg residues can be used for tryptic digestion (in bold in Figure 6B). In addition, some cleavage sites are very close together, generating peptides too small to be observed in the standard reflectron mode MALDI-TOF m/z window between 750 and 4000 Da. Instead, the two tryptic cut sites in the Pro-rich region release two large tryptic peptides, P1 and P2, which are also out of that m/z range. Despite these sequence constraints, three peptides from the *C*-terminal sequence, suitable for the standard reflectron mode MALDI-TOF MS, were found in the previously obtained spectra (Figure 6B). Their sequences were confirmed by MALDI-TOF-TOF MS/MS fragmentation (Supplementary Figure 6). In addition, *N*-terminal sequencing by Edman degradation (Irshad et al., 2008) enabled to localize the *N*-terminus of the protein downstream the *N*-terminus predicted by bioinformatics, revealing a possible *N*-terminal processing or degradation. Thus both tandem MS and Edman degradation results allowed to correctly identify At5g14920.

In order to find the P1 and P2 peptides in the Pro-rich region of At5g14920, we performed additional MALDI-TOF MS experiments, this time in linear mode instead of the reflectron mode used in PMF standard conditions. The presence of consecutive Pro in the Pro-rich region suggests that, similarly to other cell wall H/PRPs (Kieliszewski and Lamport, 1994), some of them may be hydroxylated. Up to 17 and 39 Hyp residues can be expected on P1 and P2 respectively, respectively increasing their mass up to 272 and 624 Da. This calculation excludes Hyp in the Leu-Pro and Lys-Pro contexts according to previous work (Kieliszewski and Lamport, 1994; Shimizu et al., 2005). The MS spectrum for K11 clearly shows two major peaks in the m/z range 5700–18000 (Figure 6C). The peak centered on the mean m/z 15400 possibly corresponds to P2. This m/z, which is higher than the peptide theoretical mass (between 11859 and 12483 Da depending on the number of Hyp), suggests the presence of *O*-glycosylations in this region. In addition, the width of the observed peak (m/z 14200-16600) reflects mass heterogeneity, probably due to different states of *O*-glycosylation. The di-protonated form of P2 (m/z 7700) is also observed. However, the corresponding peak in the m/z range 7000–8800 may also contain the signal of the mono-protonated form of P1 which can overlap with that of the di-protonated P2. The spectrum obtained for K14 has a peak at m/z 13070, *i.e.* within the m/z 12500–14000 range corresponding to P2. Interestingly, for this peptide a shift of mass of about 2300 Da is observed between K11 and K14, indicating a lower level of *O*-glycosylation in K14. In K14, the di-protonated form of P2 is resolved from the mono-protonated form of P1, at m/z 6500. The peak centered on m/z 7200, *i.e.* within the m/z 7000–7500 range, may be attributed to the mono-protonated form of P1. This mass range coincides with the expected mass of P1 calculated in the *N*-terminus determined by Edman degradation (up to 6866 Da, Figure 6B).

Considering the primary sequence of At5g14920 and the known codes for hydroxylation of Pro and subsequent *O*-glycosylations, At5g14920 may be *O*-gycosylated in different ways. The presence of the Thr-Pro modules suggests the presence of arabinogalactans like in AGPs (Tan et al., 2003). However, the staining of the protein with the Yariv reagent was negative (Supplementary Figure 7). The presence of Lys-Pro-Pro-Thr motifs like in the maize THRGP (threonine-hydroxyproline-rich glycoprotein) suggests the presence of arabinosylated Hyp (Kieliszewski and Lamport, 1987; Kieliszewski et al., 1990). However, similar motifs are present in AtAGP31 and in NaPRP4 which were found to contain 80 and 83% of galactose respectively (Sommer-Knudsen et al., 1996; Liu C and Mehdy, 2007). Staining of a fraction enriched in At5g14920 with a lectin specific for galactose was positive indicating the possible presence of galactose (Supplementary Figure 7). Finally, the presence of both arabinose and galactose cannot be excluded as for the *Chlamydomonas reinhardtii* GP1 HRGP (Ferris et al., 2001).

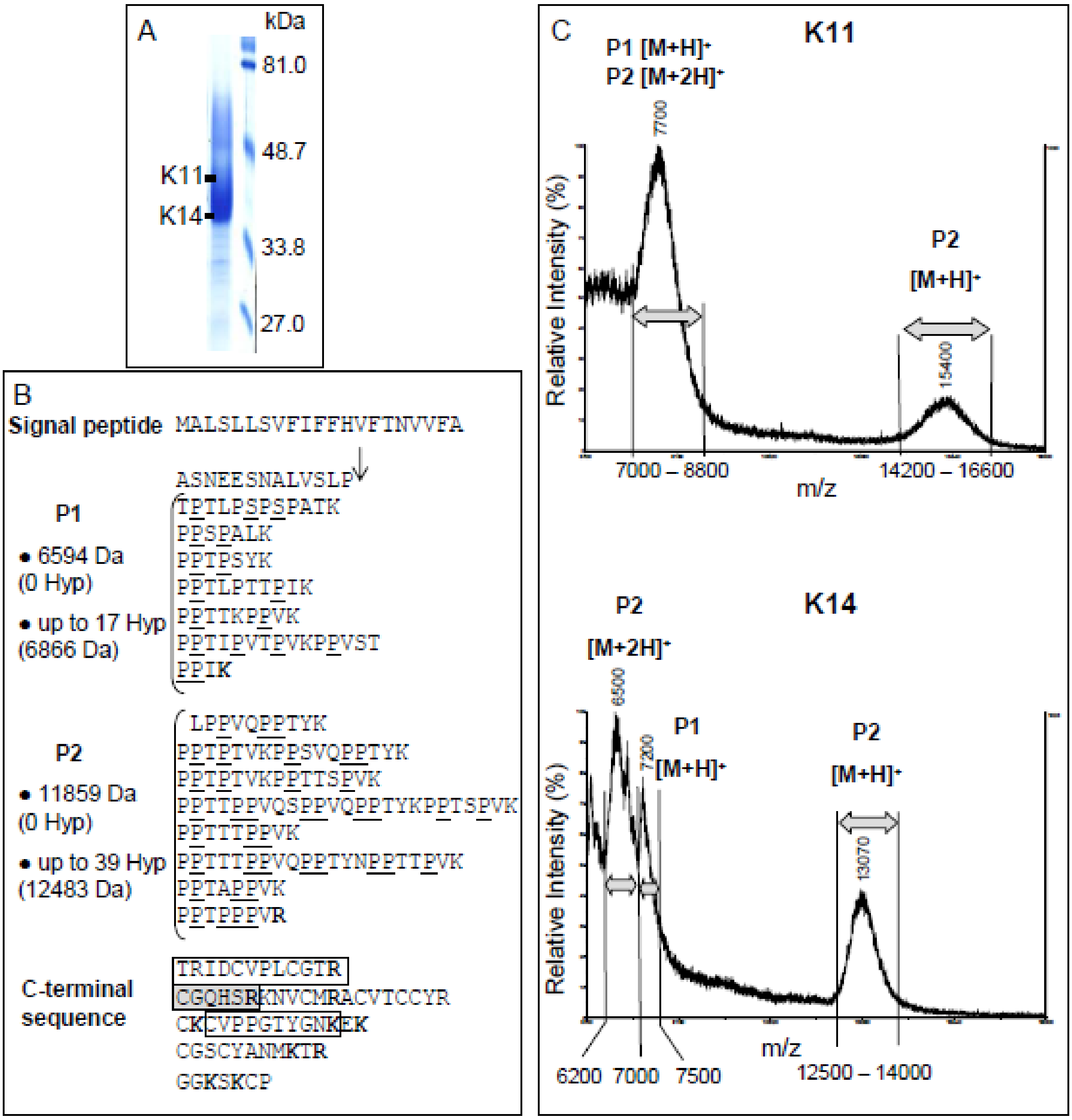

Figure 6. Identification and characterization of the At5g14920 H/PRP (NP_196996)

The protein was identified in the K (K11 to K14) CWP fraction separated by 1D-E after CEC (Irshad et al., 2008). (**A**). PMF identification was performed with the 3 *C*-terminal peptides shown in the frame on the primary sequence of the protein (**B**). The amino acid sequence of At5g14920 is written from left to right, and from top to bottom. The arrow on the primary sequence corresponds to the *N*-terminus of the mature protein as determined by Edman degradation. Lys and Arg residues sensitive to trypsin action are indicated in bold. Pro residues possibly hydroxylated are underlined. P1 and P2 resulting from tryptic digestion of the protein were observed by MALDI-TOF MS analysis in linear mode (**C**). The mass shifts observed for both peptides between K11 and K14 supposedly correspond to different states of the protein *O*-glycosylation.

The MS results reported here suggest the presence of *O*-glycosylation on an H/PRP. As for NaPRP4 and AtAGP31 (Sommer-Knudsen et al., 1996; Liu C and Mehdy, 2007), varying degrees of *O*-glycosylation were observed for At5g14920. The carbohydrate content of At5g14920 is estimated to be at much as 25 %, under that observed for NaPRP4 and AtAGP31 (75% and 80 % respectively), but much higher than that observed for soybean PRP2 (less than 1%) (Datta et al., 1989). A functional study of glycosylations should prove interesting since the role of *O*-glycosylation of H/PRPs has so far never been investigated. In the case of extensins, *O*-glycosylation was shown to keep proteins in an extended conformation (Stafstrom and Staehelin, 1986). Further work is necessary to elucidate the structure of the Pro-rich region and of its glycosylations (Kieliszewski, 2001; Schultz et al., 2004; Tan et al., 2004; Shimizu et al., 2005). However, other biochemical approaches should be considered since MS or MS/MS alone will not fully reveal the structural features of this recalcitrant protein, such as the location of Hyp residues.

Mass spectrometry as a diagnosis tool for protein structure

The various examples presented in this article illustrate that MS is a powerful tool to gain new insight into the structure of CWPs in relation to their function. However, MS does not always allow the full coverage of a protein sequence, the complete description of PTMs, and the identification of all the proteins in a complex protein mixture. It appears that some proteins or peptides present in a sample fail to be analyzed. In certain cases, the MS results should be confirmed with additional experiments. *N*- or *C*-terminal sequencing can be used to determine the location of the ends of a protein, and consequently uphold biological processes such as maturation events. To go further into the three-dimensional structure of proteins, other methods can be used such as nuclear magnetic resonance (NMR) or X-ray crystal diffraction (Liu H-L and Hsu, 2005). However, the amount of purified proteins required for these analyses can be limiting, as compared to MS. Chemical proteomics that consists in the analysis by MS of proteins specifically tagged at their site of modification, is an emerging field that offers striking insights into the functional biology of PTMs (Tate, 2008). In conclusion, MS will increasingly be used as a diagnosis tool to address biological questions given its simple use, high throughput capacity, unequaled sensitivity, and the continuous technical developments in the field.

METHODS

Plant material and cell wall isolation

Arabidopsis thaliana seeds (ecotype Columbia 0) were grown in the dark as described (Feiz et al., 2006; Irshad et al., 2008), and etiolated hypocotyls were collected at 5 and 11 days. One to 1.8 cm hypocotyls were used to isolate cell walls as previously reported (Feiz et al., 2006). The obtained cell wall fraction was ground in liquid nitrogen in a mortar with a pestle prior to lyophilization.

Cell wall protein extraction and separation

Proteins were extracted from the cell wall preparation by successive steps of 0.2 M $CaCl_2$ followed by 2 M LiCl. The proteins were separated by cation exchange chromatography (CEC). The column fractions were separated by 1D-electrophoresis (1D-E), stained and named as previously described (Irshad et al., 2008). The bands stained with Coomassie Brilliant blue were excised from the gels and digested with trypsin as previously described (Borderies et al., 2003; Boudart et al., 2005). Staining of proteins with the β-glycosyl Yariv reagent was performed as previously described (Willats and Knox, 1996). Staining of proteins with the galactose-specific peanut agglutinin (PNA) lectin was carried out using the DIG Glycan Differentiation Kit procedure (Roche, Mannheim, Germany).

Anhydrous hydrogen fluoride (HF) deglycosylation

A sample of salt extractable CWP was HF-deglycosylated for 1 h at 4 °C as described (Mort and Lamport, 1977; Shpak et al., 1999). The HF was blown off under nitrogen gas, and the deglycosylated proteins were then separated by 1D-E.

Mass spectrometry (MS) analyses

Sample preparation for all MALDI MS analyses was performed as previously described (Borderies et al., 2003). MALDI-TOF MS analyses were performed using a Voyager-DE STR mass spectrometer (Applied Biosystems/MDS, Sciex, USA). Spectra were either acquired in reflectron mode as previously reported (Borderies et al., 2003; Boudart et al., 2005) or in linear mode for a mass range of 2-20 kDa using the following parameters: accelerating voltage 25 kV, grid voltage 80 % and acceleration delay time 800 ns. Every single acquisition run was composed of 2000 laser pulses at 20 Hz. Resolution of MALDI-TOF MS in linear mode was > 300, corresponding to a full width at half maximum < 50 for a peptide of m/z 15000. The accuracy of the mass determination was ± 0.4 %, corresponding to m/z < 80 in the mass region covered in linear mode.

MALDI-TOF-TOF analyses were performed using a MALDI-TOF-TOF Voyager 4700 (Applied Biosystems/MDS, Sciex, USA). MS and MS/MS data were recorded using the following parameters: accelerating voltage 8 kV and 15 kV (source 1 and source 2 respectively) and grid voltage 86 %. The mass selection of the precursor ion was achieved using a mass window of +/- 5 Da and collision was performed in CID off mode. Data were acquired with 3750 shoots/spectrum.

Peptide mass fingerprints (PMFs) were compared to the non-redundant database of *A. thaliana* at NCBI (http://www.ncbi.nlm.nih.gov/) using ProteinProspector MS-FIT (http://prospector.ucsf.edu/cgi-bin/msform.cgi?form=msfitstandard). The criteria used for the database search were: a monoisotopic mass accuracy of 20 ppm; one missed cleavage; and hydroxylation of Pro residues. MS/MS results were analyzed using Protein Prospector MS-Product (http://prospector.ucsf.edu/cgi-bin/msform.cgi?form=msproduct).

Bioinformatics

Sub-cellular location, length of signal peptides, prediction of transmembrane domains, homologies to other proteins and protein functional domains were predicted as previously described (Minic *et al*., 2007) or using the *ProtAnnDB* database (http://www.polebio.scsv.ups-tlse.fr/ProtAnnDB/). *ProtAnnDB* collects predictions of sub-cellular location and of functional domains obtained with different software (San Clemente et al., 2009). *N*-glycosylation consensus sequences were determined using Prosite (http://www.expasy.org/prosite/). Functional domains were predicted by InterProScan (http://www.ebi.ac.uk/Tools/InterProScan/). All *A. thaliana* and *O. sativa* cell wall proteomic data are collected in the dedicated *WallProtDB* database (http://www.polebio.scsv.ups-tlse.fr/WallProtDB/).

The *ProTerNyc* software was developed in PHP5. It includes an *N*-glycan database associated with their molecular masses according to (Faye et al., 2005). From a protein sequence and a mass list, *ProTerNyc* allows (i) prediction of a signal peptide using TargetP (http://www.cbs.dtu.dk/services/TargetP/), and (ii) prediction of peptides possibly *N*-glycosylated using PROSITE (http://www.expasy.org/prosite/). In a first step, *ProTerNyc* predicts the length of the signal peptide. When the prediction is positive, a search for the presence of *N*-glycosylation motifs (PS00001) is carried out. The protein is cleaved following the prediction and submitted to MS-digest (http://prospector.ucsf.edu/cgi-bin/msform.cgi?form=msdigest). The list of masses obtained after theoretical digestion and eventually *N*-glycosylations on predicted consensus sites is compared to the list of experimental masses obtained after MALDI-TOF MS analysis with a mass tolerance up to 60 ppm. The software accepts a tolerance for the length of the signal peptide around the position predicted by TargetP. The results are given in a file that can be exported in the Microsoft Office Excel format (http://www.microsoft.com/france/office/2007/programs/excel/overview.mspx).

ACKNOWLEDGEMENTS

This study was funded by the Université Paul Sabatier in Toulouse and the Centre National de la Recherche Scientifique (France). The authors are grateful to Prof M Kielieszewski (University of Ohio, USA) for welcoming GB in her laboratory for HF deglycosylation of protein samples, and to Drs J Sampedro (University of Santiago de Compostela, Spain), Z Minic (University of Saskatchewan, Canada), and J Pelloux (University of Amiens, France) for stimulating discussions. Their gratitude also goes to G Borderies and C Pichereaux for technical assistance. All MS analyses were performed on the *Plate-forme Protéomique Toulouse Midi-Pyrénées*, France. The authors thank D Mosovich for editing the English of the text.

References

Aebersold, R., and Mann, M. (2003). Mass spectrometry-based proteomics. Nature **422,** 198-207.

Alvarez, S., Goodger, J.Q., Marsh, E.L., Chen, S., Asirvatham, V.S., and Schachtman, D.P. (2006). Characterization of the maize xylem sap proteome. J. Proteome Res. **5,** 963-972.

Baumberger, N., Doesseger, B., Guyot, R., Diet, A., Parsons, R.L., Clark, M.A., Simmons, M.P., Bedinger, P., Goff, S.A., Ringli, C., and Keller, B. (2003). Whole-genome comparison of leucine-rich repeat extensins in Arabidopsis and rice. A conserved family of cell wall proteins form a vegetative and a reproductive clade. Plant Physiol **131,** 1313-1326.

Bayer, E.M., Bottrill, A.R., Walshaw, J., Vigouroux, M., Naldrett, M.J., Thomas, C.L., and Maule, A.J. (2006). *Arabidopsis* cell wall proteome defined using multidimensional protein identification technology. Proteomics **6,** 301-311.

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

Bhushan, D., Pandey, A., Chattopadhyay, A., Choudhary, M.K., Chakraborty, S., Datta, A., and Chakraborty, N. (2006). Extracellular matrix proteome of chickpea (*Cicer arietinum* L.) illustrates pathway abundance, novel protein functions and evolutionary perspect. J. Proteome Res **5,** 1711-1720.

Birdsall, T. (1998). 5-Hydroxytryptophane: a clinically-effective serotonin precursor. Altern Med Rev **3,** 271-280.

Borderies, G., Jamet, E., Lafitte, C., Rossignol, M., Jauneau, A., Boudart, G., Monsarrat, B., Esquerré-Tugayé, M.T., Boudet, A., and Pont-Lezica, R. (2003). Proteomics of loosely bound cell wall proteins of *Arabidopsis thaliana* cell suspension cultures: A critical analysis. Electrophoresis **24,** 3421-3432.

Bosch, M., Cheung, A., and Hepler, P. (2005). Pectin methylesterase, a regulator of pollen tube growth. Plant Physiol **138,** 1334-1346.

Boudart, G., Jamet, E., Rossignol, M., Lafitte, C., Borderies, G., Jauneau, A., Esquerré-Tugayé, M.-T., and Pont-Lezica, R. (2005). Cell wall proteins in apoplastic fluids of *Arabidopsis thaliana* rosettes: Identification by mass spectrometry and bioinformatics. Proteomics **5,** 212-221.

Carpita, N.C., and Gibeaut, D.M. (1993). Structural models of primary cell walls in flowering plants: consistency of molecular structure with the physical properties of the walls during growth. Plant J **3,** 1-30.

Caspers, M., Lok, F., Sinjorgo, K., van Zeijl, M., Nielsen, K., and Cameron-Mills, V. (2001). Synthesis, processing and export of cytoplasmic endo-β-1,4-xylanase from barley aleurone during germination. Plant J **26,** 191-204.

Castellino, F., Ploplis, V., and Zhang, L. (2008). γ-glutamate and β-hydroxyaspartate in proteins. Methods Mol Biol **446,** 85-94.

Charmont, S., Jamet, E., Pont-Lezica, R., and Canut, H. (2005). Proteomic analysis of secreted proteins from *Arabidopsis thaliana* seedlings: improved recovery following removal of phenolic compounds. Phytochemistry **66,** 453-461.

Chen, X., Kim, S., Cho, W., Rim, Y., Kim, S., Kim, S., Kang, K., Park, Z., and Kim, J. (2008). Proteomics of weakly bound cell wall proteins in rice calli. J Plant Physiol **166,** 665-685.

Chevalier, F., Rofidal, V., Vanova, P., Bergoin, A., and Rossignol, M. (2004). Proteomic capacity of recent fluorescent dyes for protein staining. Phytochemistry **65,** 1499-1506.

Cho, W., Chen, X., Chu, H., Rim, Y., Kim, S., Kim, S., Kim, S.-W., Park, Z.-Y., and Kim, J.-Y. (2009). The proteomic analysis of the secretome of rice calli. Physiol Plant **135,** 331-341.

Datta, K., Schmidt, A., and Marcus, A. (1989). Characterization of two soybean repetitive proline-rich proteins and a cognate cDNA from germinated axes. Plant Cell **1,** 945-952.

Ellis, C., Karafyllidis, I., Wasternack, C., and Turner, J. (2002). The Arabidopsis mutant *cev1* links cell wall signaling to jasmonate and ethylene responses. Plant Cell **14,** 1557-1566.

Emanuelsson, O., Brunak , S., Von Heijne, G., and Nielsen, H. (2007). Locating proteins in the cell using TargetP, SignalP and related tools. Nat Protoc **2,** 953-971.

Faye, L., Boulaflous, A., Benchabane, M., Gomord, V., and Michaud, D. (2005). Protein modifications in the plant secretory pathway: current status and practical implications in molecular pharming. Vaccine **23,** 1770-1778.

Feiz, L., Irshad, M., Pont-Lezica, R.F., Canut, H., and Jamet, E. (2006). Evaluation of cell wall preparations for proteomics: a new procedure for purifying cell walls from *Arabidopsis* hypocotyls. Plant Methods **2,** 10.

Ferris, P., Woessner, J., Waffenschmidt, S., Kilz, S., Drees, J., and Goodenough, U. (2001). Glycosylated polyproline II rods with kinks as a structural motif in plant hydroxyproline-rich glycoproteins. Biochemistry **40,** 2978-2987.

Fry, S.C. (2004). Primary cell wall metabolism: tracking the careers of wall polymers in living plant cells. New Phytol **161,** 641-675.

Gray, J. (1998). Heterogeneity of glycans at each *N*-glycosylation site of horseradish peroxidase. Carbohydr Res **311,** 61-69.

Gray, J., and Montgomery, R. (2006). Asymmetric glycosylation of soybean seed coat peroxidase. Carbohydr Res **341,** 198-209.

Han, X., Aslanian, A., and Yates III, J. (2008). Mass spectrometry for proteomics. Curr Opin Chem Biol **12,** 483-490.

Hovorka, S., Biesiada, H., Williams, T., Hühmer, A., and Schöneich, C. (2002). High sensitivity of Zn^{2+} insulin to metal-catalyzed oxidation: detection of 2-oxo-histidine by tandem mass spectrometry. Pharm Res **19,** 530-537.

Irshad, M., Canut, H., Borderies, G., Pont-Lezica, R., and Jamet, E. (2008). A new picture of cell wall protein dynamics in elongating cells of *Arabidopsis*: confirmed actors and newcomers. BMC Plant Biology **8:94**.

Ito, Y., Nakanomyo, I., Motose, H., Iwamoto, K., Sawa, S., Dohmae, N., and Fukuda, H. (2006). Dodeca-CLE peptides as suppressors of plant stem differentiation. Science **313,** 8842-8845.

Jamet, E., Canut, H., Boudart, G., and Pont-Lezica, R.F. (2006). Cell wall proteins: a new insight through proteomics. Trends in Plant Sci. **11,** 33-39.

Jamet, E., Albenne, C., Boudart, G., Irshad, M., Canut, H., and Pont-Lezica, R. (2008). Recent advances in plant cell wall proteomics. Proteomics **8,** 893-908.

Jung, Y.-H., Jeong, S.-H., Kim, S., Singh, R., Lee, J.-E., Cho, Y.-S., Agrawal, G., Rakwal, R., and Jwa, N.-S. (2008). Systematic secretome analyses of rice leaf and seed callus suspension-cultured cells: Workflow development and establishment of high-density two-dimensional gel reference maps. J Proteome Res **7,** 5187-5210.

Kieliszewski, M. (2001). The latest hype on Hyp-*O*-glycosylation codes. Phytochemistry **57,** 319-323.

Kieliszewski, M., and Lamport, D. (1987). Purification and partial characterization of a hydroxyproline-rich glycoprotein in a graminaceous monocot, *Zea mays*. Plant Physiol **85,** 823-827.

Kieliszewski, M., and Lamport, D. (1994). Extensin: repetitive motifs, functional sites, post-translational codes, and phylogeny. Plant J **5,** 157-172.

Kieliszewski, M., Leykam, J., and Lamport, D. (1990). Structure of the threonine-rich extensin from *Zea mays*. Plant Physiol **92,** 316-326.

Kimura, Y., Hess, D., and Sturm, A. (1999). The N-glycans of jack bean α-mannosidase. Eur J Biochem **264,** 168-175.

Kondo, T., Sawa, S., Kinoshita, A., Mizuno, S., Kakimoto, T., Fukuda, H., and Sakagami, Y. (2006). A plant peptide encoded by *CLV3* identified by in situ MALDI-TOF MS analysis. Science **313,** 845-848.

Kurdyukov, S., Faust, A., Nawrath, C., Bär, S., Voisin, D., Efremova, N., Franke, R., Schreiber, L., Saedler, H., Métraux, J., and Yephremov, A. (2006). The epidermis-specific extracellular BODYGUARD controls cuticle development and morphogenesis in *Arabidopsis*. Plant Cell **18,** 321-339.

Lamport, D. (1965). The protein component of primary cell walls. Adv Bot Res **2,** 151-218.

Lee, R., Hrmova, M., Burton, R., Lahnstein, J., and Fincher, G. (2003). Bifunctional family 3 glycoside hydrolases from barley with α-L-arabinofuranosidase and β-D-xylosidase activity. Characterization, primary structures, and COOH-terminal processing. J Biol Chem **278,** 5377-5387.

Lige, B., Shengwu, M., and van Huystee, R. (2001). The effects of the site-directed removal of *N*-glycosylation from cationic peanut peroxidase on its function. Arch Biochem Biophys **386,** 17-24.

Lim, J.-M., Aoki, K., Angel, P., Garrison, D., King, D., Tiemeyer, M., Bergmann, C., and Wells, L. (2009). Mapping glycans onto specific N-linked glycosylation sites of *Pyrus communis* PGIP redefines the interface for EPG-PGIP interactions. J Proteome Res **8,** 673-680.

Liu, C., and Mehdy, M. (2007). A nonclassical arabinogalactan protein gene highly expressed in vascular tissues, AGP31, is transcriptionally repressed by methyl jasmonic acid in Arabidopsis. Plant Physiol **145,** 863-874.

Liu, H.-L., and Hsu, J.-P. (2005). Recent development in structural proteomics for protein structure determination. Proteomics **5,** 2056-2068.

Micheli, F. (2001). Pectin methylesterases: cell wall enzymes with important roles in plant physiology. Trends Plant Sci **6,** 414-419.

Minic, Z., Rihouey, C., Do, C., Lerouge, P., and Jouanin, L. (2004). Purification and characterization of enzymes exhibiting β-D-xylosidase activities in stem tissues of Arabidopsis. Plant Physiol **135,** 867-878.

Minic, Z., Jamet, E., Negroni, L., der Garabedian, P.A., Zivy, M., and Jouanin, L. (2007). A sub-proteome of *Arabidopsis thaliana* trapped on Concanavalin A is enriched in cell wall glycoside hydrolases. J Exp Bot **58,** 2503-2512.

Møller, I., Jensen, P., and Hansson, A. (2007). Oxidative modifications to cellular components in plants. Ann Rev Plant Biol **58,** 459-481.

Mort, A.J., and Lamport, D.T. (1977). Anhydrous hydrogen fluoride deglycosylates glycoproteins. Anal Biochem **82,** 289-309.

Oliva, M., Theiler, G., Zamocky, M., Koua, D., Margis-Pinheiro, M., Passardi, F., and Dunand, C. (2009). PeroxiBase: a powerful tool to collect and analyse peroxidase sequences from Viridiplantae. J Exp Bot **60,** 453-459.

Pattison, R., and Amtmann, A. (2009). N-glycan production in the endoplasmic reticulum of plants. Trends Plant Sci **14,** 92-99.

Roberts, K. (1990). Structures at the plant cell surface. Curr Opin Cell Biol **2,** 920-928.

Roberts, K. (2001). How the cell wall acquired a cellular context. Plant Physiol. **125,** 127-130.

Rose, J., Braam, J., Fry, S., and Nishitani, K. (2002). The XTH family of enzymes involved in xyloglucan endotransglucosylation and endohydrolysis: current perspectives and a new unifying nomenclature. Plant Cell Physiol **43,** 1421-1435.

Roudier, F., Fernandez, A.G., Fujita, M., Himmelspach, R., Borner, G.H., Schindelman, G., Song, S., Baskin, T.I., Dupree, P., Wasteneys, G.O., and Benfey, P.N. (2005). *COBRA*, an *Arabidopsis* extracellular glycosyl-phosphatidyl inositol-anchored protein, specifically controls highly anisotropic expansion through its involvement in cellulose microfibril orientation. Plant Cell **17,** 1749-1763.

Roxrud, I., Lid, S., Fletcher, J., Schmidt, E., and Opsahl-Sorteberg, H. (2007). GASA4, one of the 14-member *Arabidopsis* GASA family of small polypeptides, regulates flowering and seed development. Plant Cell Physiol **48,** 471-483.

Sampedro, J., Sieiro, C., Revilla, G., González-Villa, T., and Zarra, I. (2001). Cloning and expression pattern of a gene encoding an α-xylosidase active against xyloglucan oligosaccharides from Arabidopsis. Plant Physiol **126,** 910-920.

San Clemente, H., Pont-Lezica, R., and Jamet, E. (2009). Bioinformatics as a tool for assessing the quality of sub-cellular proteomic strategies and inferring functions of proteins: plant cell wall proteomics as a test case. Bioinform Biol Insights **3,** 15-28.

Schöneich, C. (2000). Mechanisms of metal-catalyzed oxidation of histidine to 2-oxo-histidine in peptides and proteins. J Pharm Biomed Anal **21,** 1093-1097.

Schultz, C.J., Ferguson, K.L., Lahnstein, J., and Bacic, A. (2004). Post-translational modifications of arabinogalactan-peptides of *Arabidopsis thaliana*. J. Biol. Chem. **279,** 455103-445511.

Shimizu, M., Igasaki, T., Yamada, M., Yuasa, K., Hasegawa, J., Kato, T., Tsukagoshi, H., Nakamura, K., Fukuda, H., and Matsuoka, K. (2005). Experimental determination of proline hydroxylation and hydroxyproline arabinogalactosylation motifs in secretory proteins. Plant J **42,** 877-889.

Showalter, A. (1993). Structure and function of plant cell wall proteins. Plant Cell **5,** 9-23.

Shpak, E., Leykam, J., and Kieliszewski, M. (1999). Synthetic genes for glycoprotein design and the elucidation of hydroxyproline-*O*-glycosylation codes. Proc Natl Acad Sci USA **21,** 14736-14741.

Somerville, C., Bauer, S., Brininstool, G., Facette, M., Hamann, T., Milne, J., Osborne, E., Paredez, A., Persson, S., Raab, T., Vorwerk, S., and Youngs, H. (2004). Toward a systems approach to understanding plant cell walls. Science **306,** 2206-2211.

Sommer-Knudsen, J., Clarke, A., and Bacic, A. (1996). A galactose-rich, cell-wall glycoprotein from styles of *Nicotiana alata*. Plant J **9,** 71-83.

Stafstrom, J.P., and Staehelin, L.A. (1986). The role of carbohydrate in maintaining extensin in an extended conformation. Plant Physiol **81,** 242-246.

Tan, L., Leykam, J.F., and Kieliszewski, M.J. (2003). Glycosylation motifs that direct arabinogalactan addition to arabinogalactan-proteins. Plant Physiol **132,** 1362-1369.

Tan, L., Qiu, F., Lamport, D., and Kieliszewski, M. (2004). Structure of a hydroxyproline (Hyp)-arabinogalactan polysaccharide from repetitive Ala-Hyp expressed in transgenic *Nicotiana tabacum*. J Biol Chem **279,** 13156-13165.

Tate, E. (2008). Recent advances in chemical proteomics: exploring the post-translational proteome. J Chem Biol **1,** 17-26.

Vogel, J., Raab, T., Somerville, C., and Somerville, S. (2004). Mutations in *PMR5* result in powdery mildew resistance and altered cell wall composition. Plant J **40,** 968-978.

Watson, B.S., Lei, Z., Dixon, R.A., and Sumner, L.W. (2004). Proteomics of *Medicago sativa* cell walls. Phytochemistry **65,** 1709-1720.

Willats, W.G., and Knox, J.P. (1996). A role for arabinogalactan-proteins in plant cell expansion: evidence from studies on the interaction of β-glucosyl Yariv reagent with seedlings of *Arabidopsis thaliana*. Plant J **9,** 919-925.

Willats, W.G., McCartney, L., Mackie, W., and Knox, J.P. (2001). Pectin: cell biology and prospects for functional analysis. Plant Mol Biol **47,** 9-27.

Wolf, S., Rausch, T., and Greiner, S. (2009). The N-terminal pro region mediates retention of unprocessed type-I PME in the Golgi apparatus. Plant J **58,** 361-375.

Zhu, J., Chen, S., Alvarez, S., Asirvatham, V.S., Schachtman, D.P., Wu, Y., and Sharp, R.E. (2006). Cell wall proteome in the maize primary root elongation zone. I. Extraction and identification of water-soluble and lightly ionically bound proteins. Plant Physiol. **140,** 311-325.

Supplementary data

Supplementary Figure 1. Identification of At1g78850 and At1g78860 by MALDI-TOF MS.

Supplementary Figure 2. Cartography of At1g53830 by MALDI-TOF MS.

Supplementary Figure 3. Cartography of At1g68560 by MALDI-TOF MS.

Supplementary Figure 4. MS/MS fragmentation of the P peptide of At3g16850.

Supplementary Figure 5. MS/MS fragmentation of the two modified forms of the P peptide of At3g16850.

Supplementary Figure 6. MS/MS fragmentation of a peptide of At5g14920 located at its *C*-terminus.

Supplementary Figure 7. Analysis of At5g14920 using different carbohydrate staining procedures.