Plant Soil (2006) 286:357–376
DOI 10.1007/s11104-006-9048-9

ORIGINAL PAPER

Extracellular proteomes of *Arabidopsis thaliana* and *Brassica napus* roots: analysis and comparison by MudPIT and LC-MS/MS

Urmila Basu · Jennafer L. Francis · Randy M. Whittal · Julie L. Stephens · Yang Wang · Osmar R. Zaiane · Randy Goebel · Douglas G. Muench · Allen G. Good · Gregory J. Taylor

Received: 9 March 2006 / Accepted: 22 May 2006 / Published online: 15 August 2006

Abstract An important principle of the functional organization of plant cells is the targeting of proteins to specific subcellular locations. The physical location of proteins within the apoplasm/rhizosphere at the root–soil interface positions them to play a strategic role in plant response to biotic and abiotic stress. We previously demonstrated that roots of *Triticum aestivum* and *Brassica napus* exude a large suite of proteins to the apoplasm/rhizosphere [Basu et al. (1994) Plant Physiol 106:151–158; Basu et al. (1999) Physiol Plant 106:53–61]. This study is a first step to identify low abundance extracytosolic proteins from *Arabidopsis thaliana* and *Brassica napus* roots using recent advances in the field of proteomics. A total of 16 extracytosolic proteins were identified from *B. napus* using two-dimensional gel electrophoresis, tandem mass spectrometry (LC-MS/MS) and de novo sequencing. Another high-throughput proteomics approach, Multidimensional Protein Identification Technology (Mud PIT) was used to identify 52 extracytosolic proteins from *A. thaliana*. Signal peptide cleavage sites, the presence/absence of transmembrane domains and GPI modification were determined for these proteins. Functional classification grouped the extracellular proteins into different families including glycoside hydrolases, trypsin/protease inhibitors, plastocyanin-like domains, copper–zinc superoxide dismutases, gamma-thionins, thaumatins, ubiquitins, protease inhibitor/seed storage/lipid transfer proteins, transcription factors, class III peroxidase, and plant basic secretory proteins (BSP). We have also developed an on-line, Extracytosolic Plant Proteins Database (EPPdb, http://eppdb.biology.ualberta.ca) to provide information about these extracytosolic proteins.

U. Basu (✉)
Department of Agricultural, Food and Nutritional Science, University of Alberta, 4-10 Ag/Forestry, T6G 2P5 Edmonton, AB, Canada
e-mail: ubasu@ualberta.ca

J. L. Francis · A. G. Good · G. J. Taylor · U. Basu
Department of Biological Sciences, University of Alberta, T6G 2E9 Edmonton, AB, Canada

R. M. Whittal
Department of Chemistry, University of Alberta, T6G 2G2 Edmonton, AB, Canada

J. L. Stephens
Research Services Office, University of Alberta, Edmonton, AB, Canada

Y. Wang · O. R. Zaiane · R. Goebel
Department of Computing Sciences, University of Alberta, T6G 2E8 Edmonton, AB, Canada

D. G. Muench
Department of Biological Sciences, University of Calgary, T2N 1N4 Calgary, AB, Canada

Keywords *Arabidopsis thaliana* · *Brassica napus* · Extracellular proteins · Proteomics

Abbreviations

MS	mass spectrometry
DTT	dithiothreitol
IPG	immobilized pH gradient
CHAPS	3-[(3-Cholamidopropyl) Dimethyl-Ammonio]-1-Propanesulfonate

Introduction

A remarkable diversity of micro- and macromolecular metabolites are secreted to the rhizosphere of plant roots (Bais et al. 2001, 2004). Rhizosecretion (export of compounds to the apoplasm and rhizosphere; Borisjuk et al. 1999) has been shown to be involved in processes such as nutrient acquisition, communication with other soil organisms, and resistance to disease and toxic metals (Shepherd and Davies 1993; Flores et al. 1999; Nardi et al. 2000; Walker et al. 2003). Root exudates often include phenylpropanoids and flavanoids (Walker et al. 2003), which are involved in development and interactions of roots with the environment. Low molecular weight organic molecules, mainly organic acids, amino acids, and their derivatives, play an important role in plant metal homeostasis (Basu et al. 1994; Briat and Lebrun 1999). Plant roots also secrete a battery of pathogenesis-related (PR) proteins, including β-1,3-glucanase, chitinase, and protease, to defend the plant against potential soil-borne pathogens (Bais et al. 2004). Plant roots have evolved a range of mechanisms for increasing the availability of phosphorous (P), including exudation of organic acids, and enzymes, particularly acid phosphatases (Raghothama 1999). Acid phosphatases (APases) are the most thoroughly understood root exudates (Raghothama 1999; Tomscha et al. 2004).

The mechanism by which proteins are secreted into the apoplasm/rhizosphere is not completely understood. It has been proposed that proteins are actively secreted from root epidermal cells (Flores et al. 1999; Park 2004). Several studies have suggested the possibility of vesicular trafficking and fusion as a cellular mechanism responsible for exudation (Walker et al. 2003). By generating transgenic tobacco (*Nicotiana tabacum*) expressing proteins such as the green fluorescent protein (GFP), human placental secreted alkaline phosphatase (SEAP) and xylanase in the presence of the endoplasmic reticulum (ER) signal peptide, it was shown that proteins targeted to the ER were secreted to the apoplasm, where they retained their biological activity (Gleba et al. 1999). Recombinant proteins fused to the ER-targeting signal peptide were preferentially translocated to the cell wall and extracellular space (apoplast), and subsequently secreted from the root cells (Borisjuk et al. 1999). Recently, it has been demonstrated that a functional, full-length monoclonal antibody can be secreted from transgenic *Nicotiana tabacum* roots when targeted to the endoplasmic reticulum (Drake et al. 2003). These results indicate that the ER secretory pathway is closely linked with the root secretory pathway. The involvement of membrane transporters such as the ATP-binding cassette (ABC) transporter could provide an alternative to vesicular trafficking (Jasinski et al. 2002).

Identification and characterization of extracellular proteins provides an important means of increasing our understanding of the physiological and molecular basis of plant resistance to environmental stress. The physical location of proteins within the apoplasm at the root–soil interface positions them strategically to play a role in plant response to biotic and abiotic stress. Subcellular proteomics, including subcellular fractionation, protein identification by mass spectrometry, and bioinformatics provides a powerful strategy for identification and analysis of extracellular proteins (Dreger 2003). We recently performed large-scale identification of tubulin binding proteins by LC-MS/MS after purification of proteins by tubulin affinity chromatography (Chuong et al. 2004).

In the current study we have chosen two different approaches, for the analysis of the extracellular proteins of *B. napus* and *A. thaliana*. We used 2D-PAGE, tandem mass spectrometry (LC MS/MS) and de novo sequencing for

B. napus, while the availability of a sequenced genome allowed us to take advantage of MudPIT (Multidimensional Protein Identification Technology) for *A. thaliana*. The similarity of gene sequences from *A. thaliana* and *B. napus* allowed unambiguous identification of 16 *B. napus* proteins based on homology-based searching. Since the major obstacle to identifying peptides of *B. napus* was incomplete genome sequence information, we also used *A. thaliana* as a model system for identification of extracellular proteins using MudPIT technology. MudPIT (on-line 2D LC/MS/MS system) has several advantages compared to current gel-based methods including greater peak capacity, higher sensitivity, greater throughput and a higher degree of automation (Whitelegge 2003). We successfully identified 52 proteins using MudPIT analysis of extracellular proteome of *A. thaliana*. We have also further developed an Extracytosolic Plant Proteins database (EPPdb), to provide a comprehensive source of information on extracytosolic plant proteins (Wang et al. 2004).

Materials and methods

Collection of root exudates from *Brassica napus* and *Arabidopsis thaliana*

Brassica napus cv. Westar and *Arabidopsis thaliana* ecotype Columbia (Col-0) were grown under aseptic conditions using a sterile hydroponic system described by Basu et al. (1994, 1999). This system makes use of Magenta vessels (LifeGuard^{R} Growth Vessel Product # 701010, www.osmotek.com) with built-in filters in their lids (Vented Lid Product # 750 722, www.osmotek.com, 22 mm vent). Seeds of *B. napus* were surface sterilized in 15% bleach containing Tween 20 for 15 min, followed by 2 min in 70% ethanol, and three rinses in sterile water. Approximately 30 seeds were germinated per plate on seed germination media (2.2 g MS salt mixture, 10 g sucrose, 10 g agar l^{-1}, pH 5.8) for 2–3 days. Seedlings were then transferred to Magenta vessels containing liquid media (1 mM $Ca(NO_3)_2$, 300 μM $Mg(NO_3)_2$, 300 μM NH_4NO_3, 400 μM KNO_3, 100 μM K_2HPO_4, 100 μM K_2SO_4, 6 μM H_3BO_3, 2 μM $MnCl_2$, 0.15 μM $CuSO_4$, 0.5 μM $ZnSO_4$, 10 μM $FeCl_3$, 10 μM Na_2EDTA, pH 5.5 with MES buffer). A small number of plants per container (~ 4 plants in 60 ml of growth solution) were used to minimize risk of contamination and to control plant-induced pH changes in the growth solution. To avoid mixing of seed proteins with root extracellular proteins, seed coats were removed prior to transferring seedlings from agar plates into liquid medium and seedlings were grown in this media for 4 days. Seedlings were transferred and grown for an additional 4 days in fresh liquid media, which served as the collection media for root extracellular proteins. Shaking ensured that plant roots received adequate aeration throughout the growth period. At the end of the experimental period, individual containers were checked for sterility and bulked to provide sufficient protein for analysis (80–100 μg).

Seeds of *A. thaliana* ecotype Columbia were surface sterilized in 30% bleach containing Tween 20 for 10 min and rinsed in sterile water five times. The seeds were suspended in 0.1% Bacto Agar and approximately 15 seeds were plated onto a photographic slide containing fine mesh on a seed germination plate (2.2 g MS Basal medium, 15 g sucrose, 7 g phytagar L^{-1}, pH 6.0). After 3 days of cold incubation (4°C), plates were transferred to the growth chamber for 8 days. For collection of extracellular proteins, a sterile hydroponic system was used as described above. The slides containing the seedlings were transferred to vessels containing *Arabidopsis* maintenance media (0.5 mM $Ca(NO_3)_2$, 0.5 mM $MgSO_4$, 1.25 mM KNO_3, 0.625 mM KH_2PO_4, 1.75 μM H_3BO_3, 3.5 μM $MnCl_2$, 0.125 μM $CuSO_4$, 0.25 μM $ZnSO_4$, 2.5 μM NaCl, 0.025 μM $CoCl_2$, 3.125 μM $FeCl_3$, 3.125 μM Na_2EDTA at pH 6.0) modified from Richards et al. (1998) for 1 day. The seedlings were grown for an additional 10 days in fresh maintenance media, which served as the collection media.

Sample preparation for 2D gel electrophoresis

Sterile solutions were concentrated at 4°C using a pressure ultrafiltration system (Amicon 8200) with YM3 membrane (3 kD exclusion limit) to

retain proteins above 3 kD and remove salts in the ultrafiltrate. The proteins were further concentrated using a Speed Vac concentrator and stored at –20°C until further analysis. The concentrated *B. napus* protein sample was passed through BioGel P6 to remove salts and other low molecular weight compounds below 6 kD. *Arabidopsis thaliana* exudates were cleaned using the ReadyPrep 2-D-clean up kit (BioRad, Hercules, CA, USA) before MudPIT analysis. After protein quantitation using the Bradford method (Bradford 1976), samples (8–12 μg protein) were suspended in sample rehydration buffer (8 M Urea, 0.2% Carrier ampholytes, 50 mM DTT, 4% CHAPS and 0.0002% bromophenol blue). Samples were loaded (overnight) into 7 cm, pH 3–6 or pH 4–7 immobilized pH gradient (IPG) gel strips (BioRad, Hercules, CA, USA) by the in-gel rehydration method. Proteins were subsequently focused for 30 min at 250 V, 2.5 h on a linear gradient from 250 V to 4,000 V and 2.5 h at 4,000 V at 20°C; current limited to 50 μA/strip. Strips not immediately processed after the first dimension were stored at –70°C.

Before transfer to the second dimension, strips were incubated for 10 min in 2 ml equilibration buffer 1 (6 M Urea, 0.05 M Tris pH 8.8, 2% SDS, 20% glycerol, 2% (w/v) DTT) followed by 10 min incubation in 2 ml equilibration buffer 2 (6 M Urea, 0.05 M Tris pH 8.8, 2% SDS, 20% glycerol, 2.5% (w/v) iodoacetamide). The IPG strips were loaded on top of lab-cast, 1 mm thick, SDS polyacrylamide gels (12.5% or 15%) and run at a constant voltage of 160 V until the dye front reached the gel border. Proteins were visualized by silver staining according to Shevchenko et al. (1996). Gels were scanned and converted to digital images using an Alpha Innotech Imaging System (Alpha Innotech Corporation, San Leandro, CA, USA).

Tandem mass spectrometry for analysis of *B. napus* extracellular proteins

Protein spots were excised manually from gels using a sterile pipette tip in a laminar flow hood and stored at –80°C until subsequent analyses. Individual or pooled gel spots (2–3 spots) from silver stained gels were subjected to robot-controlled automated gel digestion using trypsin (Promega sequencing grade modified) and the resulting peptides were extracted from the gel. An arbitrary, 3 mm^2, rectangular region from a part of the gel where there was no visible protein spots was used as a control. The tandem MS experiments were performed by analyzing tryptic peptide mixtures with an automated, on-line, capillary LC HPLC (Waters, USA) coupled to a Q-TOF MS/MS system (Micromass, UK). Tryptic peptides were separated using a linear water/acetonitrile gradient (0.2% Formic acid) on a Picofrit reversed-phase capillary column, (5 μm BioBasic C18, 300°A, 75 μm ID × 10 cm, 15 μm tip) (New Objectives, MA, USA), with an in-line PepMap column (C18, 300 μm ID × 5 mm; LC Packings, CA, USA) used as a loading/desalting column. Electrosprayed samples were scanned from 400–1,600 m/z and MS/MS scans were collected from 50–2,000 m/z. The resulting tandem mass spectra were used for de novo peptide sequence determination.

MudPIT (Multi-dimensional reverse-phase chromatography with on-line tandem mass spectrometry) for analysis of *A. thaliana* extracellular proteins

Arabidposis thaliana extracellular proteins were analyzed at the University of Victoria, British Columbia Proteomics Centre. Protein samples (150 μg) were digested with 10 μg of Porcine trypsin (Promega), dissolved in 5% acetonitrile with 3% Formic acid, and loaded onto a strong cation exchange column (500 μm ID × 15 mm BioX-SCX 5 μm, connected to Valve B of the SwitchOS) and then gradually released to a reverse phase column (300 μm ID × 1 mm PepMap C18, 5 μm, 100°A nano precolumn LCPackings/Dionex, connected to Valve A of the SwitchOS) by stepwise elution with salt steps of increasing molarity. The elute from this column was allowed to divert to waste for 4 min using the SwitchOSII, then the flowpath was diverted to the Ultimate pumps and the sample was eluted onto a 75 μm ID × 15 cm PepMap C18 3 μm, 100°A nanocolumn (LCPakings/Dionex LC Packings, San Francisco, CA). The column was sleeved via 20 cm of 20 μm ID fused silica (PolyMicroTechnologies) to a Valco stainless steel zero dead

volume fitting which had a high voltage lead (2,500 V) and a 10 μm New Objective fused silica tip emitter (PicoTip™ New Objective, Woburn, MA) positioned at the orifice of an Applied Biosystems/Sciex QStar Pulsar I Quadrupole time-of-flight mass spectrometer coupled to a Protana nanospray source (Proxeon, Denmark). All MS analyses were performed on an MDS SCIEX API QSTAR Pulsar in positive ion mode (PE SCIEX Concord, Ontario, Canada).

The mass spectrometer independent data acquisition parameters were as follows: after a 1 s survey scan from 300–1,500 m/z peaks with signal intensity over 10 counts with charge state 2–5 were selected for MS/MS fragmentation using software determined collision energy. Next a two second MS/MS from 65–1,800 m/z was collected for the three most intense ions in the survey scan. Further peptides were eluted from the SCX column using 50 μl volume of 25, 50, 75, 100, 150, 200, 250, 500, 1000, 2000 mmol ammonium acetate pH 4.0 gradient and the above gradient was performed for each salt injection.

Data analysis

For *A. thaliana*, the software used in the analysis was ProID and the NCBI non-redundant database was searched with an error tolerance of 0.15 Da for both the MS and the MS/MS scans. ProID enables the rapid identification of proteins from LC/MS/MS data files. Every MS/MS spectrum in the data file is used to search a protein or DNA sequence database using the Interrogator™ Search algorithm. Interrogator Search is a database-searching algorithm that uses fragment ion masses to determine the identity of the peptide. The results were then written by the software to a Microsoft Access database. The resulting database was then queried using a minimum confidence limit of 50 and a protein score of 15 and the oxidation of methionine was selected as a variable modification. When multiple peptides were identified the score and best confidence are reported for the peptide with the highest values.

For *B. napus*, protein identification from the generated MS/MS data was first attempted by searching the NCBI non-redundant database using Mascot Daemon (Matrix Science, UK). Search parameters included carbamidomethylation of cysteine, possible oxidation of methionine and one missed cleavage per peptide. When necessary, the acquired MS/MS data were further processed using MassLynx 4.0 software. De novo peptide sequence determination was completed. Database searching was carried out by submitting the predicted peptides to MS-Blast (http://dove.embl-heidelberg.de/Blast2/msblast.html) as well as to MS-Homology (http://prospector.ucsf.edu/ucsfhtml4.0/mshomology.html), to determine a match to a homologous protein in other plant species. For each protein spot, the candidate ranked at the top of the list was considered a probable positive identification. The presence of putative signal peptides was predicted using the TargetP Server v1.01 program (Emanuelsson et al. 2000). The amino acid sequences of hypothetical proteins were also analyzed by the TMHMM prediction method for transmembrane helices (Krogh et al. 2001). The Glycosylphosphatidylinositol (GPI) Modification Site Prediction was done using the server http://mendel.imp.ac.at/gpi/plant_server.html (Eisenhaber et al. 2003).

Development of an Extracytosolic Plant Protein Database (EPPdb)

We have developed an on-line Extracytosolic Plant Proteins Database (EPPdb; http://eppdb.biology.ualberta.ca) providing information about the extracytosolic proteins to the plant biology committee (Wang et al. 2004). For *B. napus* proteins, the individual protein entries are hyperlinked to the relevant spots on an image map created from the reference gel.

Results and discussion

Two-dimensional gel electrophoresis of extracellular root proteins from *B. napus* and *A. thaliana*

We optimized a sterile hydroponic system for collection of root exudates from *B. napus* using growth conditions adapted from protocols previously established for wheat (Basu et al. 1994, 1999). Changing the type of the lids (larger vent

opening) resulted in healthier plants. Two-dimensional gels of the extracellular plant proteins of *B. napus* followed by silver staining showed approximately 20–25 proteins spots (Fig. 1). At least 4 gels were run for every analysis and the profile was very consistent. Close examination of these gels indicated that there were an additional 25–30 less abundant proteins. Two-dimensional gel profiles were consistent between independently collected batches, showing only differences in spot intensity rather than in spot number or position (data not shown). These quantitative differences were likely the result of variation in the amount of protein loaded onto the gels between experiments and staining differences. Figure 2 displays a silver-stained, two-dimensional map of *A. thaliana* extracellular root proteins separated by pH 3.0–6.0 IEF/15% gel containing 20–25 abundant proteins and 30–35 less abundant proteins. Overall, the protein spots were clearly resolved and the 2-D gel patterns were highly reproducible (data not shown).

LC-MS/MS analysis of *B. napus* extracellular plant proteins

In the most common bottom-up approach (MALDI-TOF), proteomic studies rely on separation of proteins by 2-dimensional gel electrophoresis, excision of individual protein spots from the gel, cleavage reactions, extraction of peptides, mass spectrometry and identification of proteins by database searches (Pandey and Mann 2000). Since our major obstacle to identifying peptides was incomplete sequence information available for the *Brassica* genome, peptide sequence analysis by tandem mass spectrometry, on-line liquid chromatography coupled to Q-TOF MS/MS was used for unambiguous identification of proteins. The tandem mass spectrometer was used to sequence peptides by generating fragments via collision-induced dissociation, with subsequent measurement of the masses of the fragments. The resulting MS/MS spectrum was compared to a theoretically generated spectrum from a database of known proteins and ranked by score. However, for most proteins there was no match found by searching against a database of theoretical sequences; thus de novo sequence analysis was the only way to confirm the depicted sequences and identify proteins. In addition to being a method of primary peptide identification, de novo sequencing algorithms can also be used to simply filter out low-quality spectra. For determining peptide sequences via de novo sequencing, all spectra were interpreted manually using known rules and established principles (Medzihradszky and Burlingam 1994).

Bioinformatic analyses of the sequences derived from de novo peptide sequence determination were then performed against the database using the MS-Homology and MS-Blast programs.

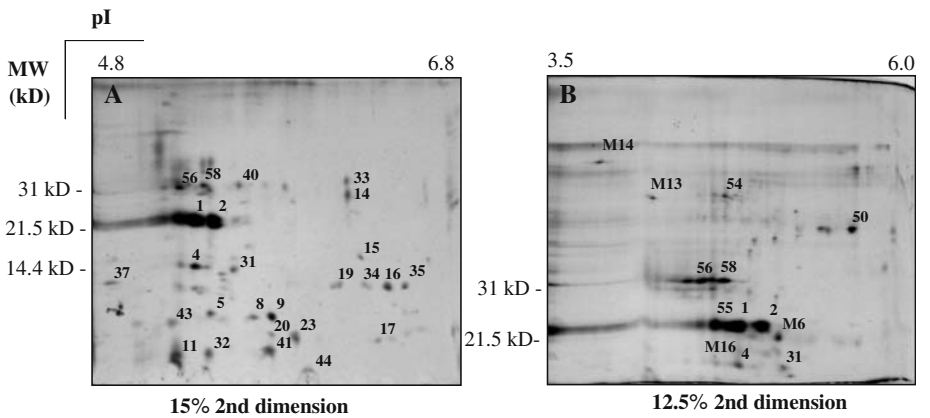

Fig. 1 Two-dimensional maps showing separation of *B. napus* extracytosolic root proteins by pH 4.0–7.0 IEF/15% gel (A) and pH 3.0–6.0 IEF/12.5% gel (B). Gels were stained with silver nitrate. The resulting images were cropped to focus in on areas where most of the spots were visible

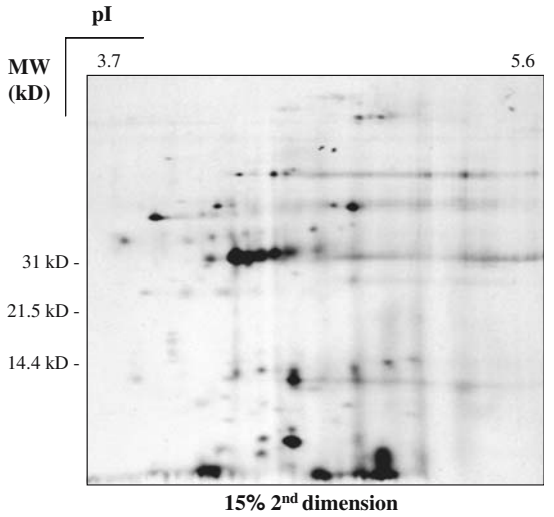

Fig. 2 Two-dimensional map showing separation of *A. thaliana* extracytosolic root proteins by pH 3.0–6.0/ 15% gel. Gel was stained with silver nitrate. The resulting image was cropped to focus in on areas where most of the spots were visible

An example for one of the peptides (precursor ion MH^+ 722.38 from spot 1 of Fig. 1) is shown in Fig. 3. The high confidence limit settings that were used in the analysis of the peptide data coupled with the identification of multiple peptides for most of the proteins, allowed for the unambiguous identification of individual proteins using this technique.

In total, 25 of the 50–60 *B. napus* protein spots resolved by 2-D gel electrophoresis were excised from gels, and analyzed by tandem mass spectrometry. These spots were chosen because they were well resolved and more abundant when visualized with silver staining. The identities of 16 of these extracellular proteins (and the matching sequences) are listed in Table 1. The experimentally determined isoelectric points and molecular masses of the proteins were generally consistent with the predicted molecular masses and isoelectric points of the corresponding proteins from the database (Table 1). The discrepancy between the experimental and theoretical values observed for some proteins might be explained by proteolytic degradation of polypeptides during sample preparation, post-translational modification events, or variability arising from alternate splicing of mRNAs. Other studies have observed similar levels of discrepancy between the predicted and experimental molecular masses and isoelectric points of proteins identified by mass spectrometry (Chang et al. 2000). Three of our most abundant root extracellular proteins (spot 1, MW 23 kD, pI 5.0; spot 2, MW 23 kD, pI 5.2 and spot M6, MW 21 kD, pI 5.25 from Fig 1) have been identified as putative trypsin inhibitors (Table 1). Protein spots 1 and 2 migrated in the gel with a similar molecular mass but with a different isoelectric point (Fig 1). Protein spots 56 and 58, which also exhibited similar molecular weight to each other and different isoelectric points, were identified as endochitinases (Fig 1B). Spots 8 and 14 (Fig 1A) that have different molecular weights and isoelectric points matched to the same Stellacyanin (uclacyanin 3)-like protein (Q9LY37). These modifications suggest that these proteins are either closely related isoforms of the same protein or are identical proteins with different post-translational modifications.

MudPIT analysis of *A. thaliana* extracellular plant proteins

While two-dimensional gel electrophoresis is a powerful technique for protein separation, it has a number of inherent limitations. Consequently we studied the extracellular proteome of *A. thaliana* by using MudPIT technology. Multidimensional chromatography coupled to tandem mass spectrometry (LC/LC-MS/MS) represents a promising alternative for proteome analysis (Peng et al. 2003; Whitelegge 2003). We used an online approach for 2D chromatography, where an extracellular protein mixture from *A. thaliana* roots was applied to an SCX chromatography column and discrete fractions of the adsorbed peptides were subsequently displaced directly onto the RP chromatography column using a salt step gradient. Peptides were then eluted and analyzed by tandem mass spectrometry. Both the MS and MS/MS scans were further analyzed by searching against the NCBI non-redundant database. The resulting database was then queried with a minimum confidence limit of 50 and a score of 15 using this approach. We have obtained the identity of 52 *A. thaliana* extracellular proteins (Table 2).

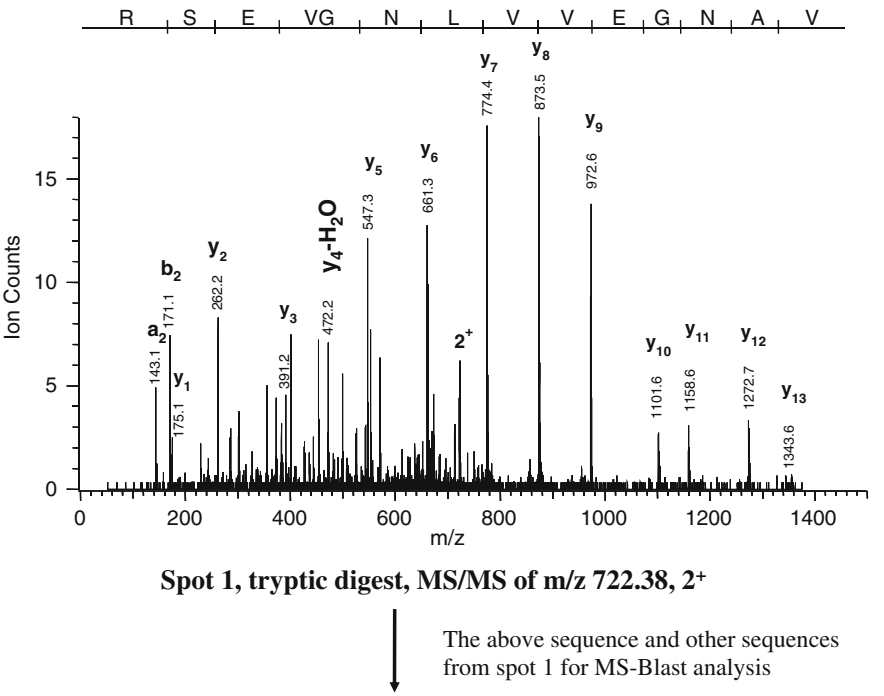

Fig. 3 Collision-induced dissociation spectrum of the *m*/*z* peptide 722.38 derived from spot 1. A partial MS/MS spectrum of this peak was used for de novo peptide sequence determination and is shown with the assignment of some ions, leading to sequence information. The *roman series* (*b* and *y*) corresponds to single-charged fragment ions. The predicted peptide was then submitted to MS-BLAST to find a match to a homologous protein in other plant species. Unmatched sequences are indicated in bold

Band 1 MS-BLAST Assignment: sptrembl|Q9M8Y9|Q9M8Y9 Putative Trypsin Inhibitor [*Arabidopsis thaliana*] Total Score: 201

Query:	BFANPSQCGESGVWR	VANGEVVLNGVESR	CPHQPVM	PVMF
Sbjct:	**KFVDPRP**CGESG**F**WR	**SSE**GEVVLNG**S**ES**T**	CP**Q**QP**L**M	PVMF

Target P and TMHMM analyses of extracellular plant proteins

Most extracellular proteins possess various distinct features responsible for their translocation from the cytoplasm into the extracellular environment. One of these features is the presence of a cleavable signal peptide on the N-terminus that is responsible for targeting them to ER, the first organelle in the secretory pathway (Vitale and Denecke 1999). The signal peptide does not possess a consensus amino acid sequence, but is characterized by three conserved domains; the positively charged n-region on the N-terminus, and a central hydrophobic h-region followed by a polar c-region containing the cleavage site. Another feature of extracellular proteins (and all soluble or non-membrane proteins) is the absence of transmembrane domains in their amino acid sequences. In the secretory pathway, proteins leave the ER for the Golgi complex where they are packaged into vesicles destined for the plasma membrane, the site of secretion to the extracellular matrix (Vitale and Denecke 1999; Crofts et al. 1999). If an extensive hydrophobic domain is present in the protein, then it gets embedded in the plasma membrane, and the absence of such a domain would result in secretion to the extracellular matrix.

We continued our bioinformatic analysis by identifying signal peptides and determining the presence/absence of transmembrane domains of these extracytosolic proteins. The plant predictor version of TargetP, which predicts the subcellular location of eukaryotic protein sequences (Emanuelsson et al. 2000) was used for the analysis of signal peptides of extracytosolic proteins. The first 130 residues (size recommended in TargetP instructions) from the N-terminus of each of the top hits were submitted to TargetP.

Table 1 Identification of proteins from the *Brassica napus* extracellular proteome using LC-MS/MS and homology based searching

Spot	Accession/ Matched protein	Peptide Sequence (m/z)	Matched sequence	Expected size (kD) /pI	Calculated size (kD) /pI	Target P	TMHMM*	GPI
1	Q9M8Y9 Putative trypsin inhibitor – *A. thaliana*	FANPSKCGESGVWR (1594.71) VANGEVVLNGVESR (1442.77) CPHQPVMF (1116.49) SCKGSLSWETGAAEGN (1653.71) LLPSSTV (1399.78)	F**V**D**P**R**P**CGESG**F**WR **SSE**GEVVLNG**S**ES**T** CP**Q**QP**L**M**V**	22.9/5.9	23/5.0	S (1) 21	0	None
2	Q9M8Y9 Putative trypsin inhibitor – *A. thaliana*	FANPSKCGESGVWR (1594.71) VANGEVVLNGVESR (1442.77) CPHQPVMF (1116.49) SCKGSLSWETGAAEGN (1653.71) LLPSSTV (1399.78)	F**V**D**P**R**P**CGESG**F**WR **SSE**GEVVLNG**S**ES**T** CP**Q**QP**L**M**V**	22.9/5.9	23/5.2	S (1) 21	0	None
4	Unknown, a putative glycoprotein	YVVSLLDEK VNFEYCNK FVLPLGEVPK			17/5.0	N/A	N/A	None
8	O9LY37 Stellacyanin (uclacyanin 3)-like protein – *A. thaliana*	VGDTLEFK (908.47)	VGDTLEFK	19.38/5.5	10.5/5.3	S (1) 23	1 (4–26*)	164, 163
9	Q96316 Blue copper-binding protein III (uclacyanin 3) – *A. thaliana*	VGDILEFK (920.55) AGYDNCDSSAAT (2200.82)	VGD**T**LEF**V** AGYDNCDSS**G**AT	22.52/5.05	10.5/5.5	S (1) 21	2 (5–27*, 204–221)	198
12	Q9ZV18 Putative protease inhibitor	VM(AD)VLLDGTPVT [GD/AT]FR (1804.90)	V**N**A**A**V**I**LDG**S**PVT**A**DFR	7.62/6.14	8/5.3	_(2)-	0	None
14	Q9LY37 Stellacyanin (uclacyanin 3)-like protein – *A. thaliana*	VGDTLEFK (908.47)	VGDTLEFK	19.38/5.5	29.5/5.9	S (1) 23	1 (4–26*)	164, 163
16	O81352 Cytosolic Cu/Zn superoxide dismutase *Brassica rapa*	GVAVLNSSEGVK (1159.63)	GVAVLNSSEGVK	15.17/5.64	14/6.0	_(3) -	0	None
19	Q9FK60 Cu/Zn superoxide dismutase-like protein – *A. thaliana*	AVVVHADPDDLGK (1334.74)	AVVVHADPDDLGK	16.94/7.16	14/5.8	_(5) -	0	None
44	P59263 Ubiquitin – *A. thaliana*	TLADYNIQK (1064.56)	TLADYNIQK	8.53/6.56	8.5/5.6	_(2)-	0	None

Table 1 continued

Spot	Accession/ Matched protein	Peptide Sequence (m/z)	Matched sequence	Expected size (kD) /pI	Calculated size (kD) /pI	Target P	TMHMM*	GPI
50	P24102 Peroxidase 22 [Precursor] – *A. thaliana*	(AQ)CQFVTPR (1105.58)	AQCQFVTPR	38.11/5.66	38/5.5	S (1) 29	0	None
54	Q8VXZ7 Putative alpha-galactosidase – *A. thaliana*	APLLLGCDVR (1113.62) (DL)LAVNQDPLGVQGR (1594.83)	APLL**I**GCDVR **EII**AVNQDPLGVQGR	48.36/4.73	48/4.7	S (1) 30	1 (12–31*)	None
56	Q06209 Basic endochitinase CHB4 [Precursor] – *Brassica napus*	PVLSQGFGATLR (1244.60) TALWFWVNNVR (1404.68) VSFLNAANTFPSFANSVSR (2027.94)	PVL**N**QGFGAT**I**R T**G**LWFW**M**N**S**VR **D**SF**I**NAANTFP**N**FANSV**T**R	28.7/8.28	28/4.6	S (1) 24	0	None
58	Q06209 Basic endochitinase CHB4 [Precursor] – *Brassica napus*	PVLSQGFGATLR (1244.60) TALWFWVNNVR (1404.68) VSFLNAANTFPSFANSVSR (2027.94) LELATMFAH (1031.48) TFPSFANSVSR (1211.62) LNAANTFPSFANSVSR (1694.76)	PVL**N**QGFGATLR TALWFW**M**N**S**VR **D**SFLNAANTFP**N**FANSV**T**R **R**E**I**ATMFAH TFP**N**FANSV**T**R INAANTFP**N**FANSV**T**R	28.7/8.28	28/4.9	S (1) 24	0	None
M6	Q9M8Y9 Putative trypsin inhibitor– *A. thaliana*	(VA)N(GE)VVLN(GV)ESR (1441.75)	**SS**EGEVVLNG**S**ES**T**	22.9/5.9	21/5.25	S (1) 21	0	None
M13	Q8VXZ7 Putative alpha-galactosidase– *A. thaliana*	APLLIGCDVR (1112.61) EIIAVNQDPLGVQGR (1593.88) YDNCFNLGIKPIER (1737.89)	APLLIGCDVR E**I**IAVNQDPLGVQGR YDNCFNLGIKPIER	48.36/4.73	48/4.7	S (1) 30	1 (12–31*)	None
M19	Q9ZUJ8 Plant Basic Secretory Protein (BSP) family protein	WDQGYDVTAR (1209.58) FLEYCNDLR (1214.61) VDFSVVDNTGDSPGGR (1594.79)	WDQGYDVTAR FL**E**YCNDLR VDFSVVDNTGDS**P**GGR	25.4/5.95	25/5.35	S (1) 21	1 (7–27*)	None

Unmatched sequences are indicated in bold. Amino acids with in brackets indicate that order of these two amino acids could not be determined by mass spectrometry. Masses are monoisotopic

Q9ZUJ8 is predicted to contain 1 TM domain (7–27 of 225) that is likely a signal peptide (SP RC = 1). Q9LY37 is predicted to contain 1 TM domain (4–26 of 187) that is likely a signal peptide (SP RC = 1). Q96316 is predicted to contain 2 TM domains. The first (5–27 of 222) is likely a signal peptide (SP RC = 1). The second (204–221 of 222) might be a GPI anchor modification site (198 predicted as potential GPI modification site). Q8VXZ7 is predicted to contain 1 TM domain (12–31 of 437) that is likely a signal peptide (SP RC = 1)

Table 2 Identification of proteins from the extracellular proteome of *Arabidopsis thaliana* using MudPIT

Accession/ AGI numbers/ Matched protein	Peptide sequence	Matched sequence	Number of distinct peptides	Best confidence	Score	Target P	TMHMM*	GPI†
O81862, At4g19810 Putative chitinase	AVLGFPYYGYAWR LTNANSHSYYAPTT-GAAISPDGSIGYGQIR SAVVAEASSSGKPR SWIQAGLPAKK TAYASMASNPTSR VTGPPAALFDPSNAGPSGD AGTR	AVLGFPYYGYAWR LTNANSHSYYAPTT-GAAISPDGSIGYGQIR SAVVAEASSSGKPR SWIQAGLPAKK TAYASMASNPTSR VTGPPAALFDPSNAGPSGD AGTR	7	99	35	S(1) 24	0	None
P28493, At1g75040 Pathogenesis-related protein 5 [Precursor]	FNTDQYCCR GANDKPETCPPTDYSR LGDGGFELTPGASR NNCPTTVWAGTLAGQGPK QLTAPAGWSGR YAGJVSDLNAACPDMLK + Oxidation (M)	FNTDQYCCR GANDKPETCPPTDYSR LGDGGFELTPGASR NNCPTTVWAGTLAGQGPK QLTAPAGWSGR YAG**C**VSDLNAACPDMLK	6	99	33	S (1) 20	1 (7–24*)	None
Q8GWI7, At1g52070 Hypothetical protein (Putative jasmonate inducible protein)	ISPTYGFVSGTK MVIGFHGSAGK QWDDGADHDNVAK SINVDYEK TSEVIGYPK	ISPTYGFVSGTK MVIGFHGSAGK QWDDGADHDNVAK SINVDYEK TSEVIGYPK	5	99	27	S (1) 17	0	None
Q940G9, At5g09530 Periaxin-like protein.	IPEVPKPELPK OPEIPKPELPK VPEIPKPELPK VPEIQKPELPK VPEVPKPELPTVPEVPK	IPEVPKPELPK **F**PEIPKPELPK VPEIPKPELPK VPEIQKPELPK VPEVPKPELPTVPEVPK	5	99	26	S (1) 33	1 (7–26*)	None
P33154, At2g14610 Pathogenesis-related protein 1 precursor (PR-1)	ANYNYAANTCNGV-CGHYTQVVWR GAVGVGPMQWDER + Oxidation (M) CNNGGTIISCNYDPR	ANY-NYAANTCNGVC-GHYTQVVWR GAVGVGPMQWDER CNNGGTIISCNYDPR	4	99	29	S (1) 26	0	None
O23248, – Class IV chitinase	AINGALEJDGA-NTATVQAR DYCDENATQYPCNPNK TALWYWTNR VQPVISQGFGATIR	AIN-GALE**C**DGANTATVQAR DYCDENATQYPCNPNK TALWYWTNR VQPVISQGFGATIR	4	99	26	S (1) 28	0	None

Table 2 continued

Accession/ AGI numbers/ Matched protein	Peptide sequence	Matched sequence	Number of distinct peptides	Best confidence	Score	Target P	TMHMM*	GPI†
O65351, At5g67360 CUCUMISIN-like serine protease (Putative erine protease)	HVVGSPVAISWT ISVEPAVLNFK SVHPEWSPAAIR TVTSVGGAGTYSVK	HVVGSPVAISWT ISVEPAVLNFK SVHPEWSPAAIR TVTSVGGAGTYSVK	4	99	28	S (1) 24	0	None
Q9ZV19, At2g38860 Putative protease inhibitor	EN- PTVNAAVILDGSPVTADFR KNSWPELTGTNGDYAAV- VIER VFVDGNR	EN- PTVNAAVILDGSPVTADFR KNSWPELTGTNGDYAAV- VIER VFVDGNR	4	99	31	_ (2) -	0	None
NP_567291.1, At4g05320 Polyubiquitin	TLADYNIQKESTLHLVLR ESTLHLVLR TLADYNIQK	TLADYNIQKESTLHLVLR ESTLHLVLR TLADYNIQK	3	99	28	_ (2) -	0	None
AAM63339.1, At3g57260 β-1, 3-glucanase2 (BG2) (PR-2)	LASSQTEADKWVQENVQ- SYR TYVNNLIQHVK VSTAIATDTTTDTSPPSQGR	LASSQTEADKWVQENVQ- SYR TYVNNLIQHVK VSTAIATDTTTDTSPPSQGR	3	99	32	S (1)30	1(7–29*)	None
O80994, At2g26020 Probable cysteine-rich antifungal protein	HGSCNYVFPAHK LCEKPSGTWSGVCGN- SNACK NQCINLEGAK	HGSCNYVFPAHK LCEKPSGTWSGVCGN- SNACK NQCINLEGAK	3	99	30	S (1) 29	1 (4–21*)	None
Q9ZU19, At1g52060 F5F19.12 protein	ALNSIGAYLTK ISPTYGVVSGTK TSEVIGYPK	ALNSIGAYLTK ISPTYGVVSGTK TSEVIGYPK	3	99	26	_ (1) -	0	None
AAA32765.1, At2g27030 Calmodulin 3	DTDSEEELKEAFR VFDKDQNGFISAAELR	DTDSEEELKEAFR VFDKDQNGFISAAELR	2	99	27	_ (2) -	0	None
AAF99769.1, At1g08830 F22013.32	AV- VVHADPDDLGKGGHEL- SLATGNAGGR	AV- VVHADPDDLGKGGHEL- SLATGNAGGR	2	99	34	_ (3) -	0	None
AAN60254.1, -Unknown	ATYHFYNPAQNNWDLR AVSAYCSTWDADKPYAWR	ATYHFYNPAQNNWDLR AVSAYCSTWDADKPYAWR	2	99	24	S (3) 21	0	None
1JXCA, -Trypsin Chymotrypsin inhibitor	EYGGDVGFGFCAPR IFPTICYTR	EYGGDVGFGFCAPR IFPTICYTR	2	99	22	_ (2) -	0	None
NP_175617.2, At1g52050 Jacalin lectin family protein	KVYVTFTETHIR	KVYVTFTETHIR	2	99	29	S (3) 17	0	None

Table 2 continued

Accession/ AGI numbers/ Matched protein	Peptide sequence	Matched sequence	Number of distinct peptides	Best confidence	Score	Target P	TMHMM*	GPI†
NP_564067.1, At1g18970 Germin-like protein (GLP1) (GLP4)	GLVHFQWNVGQVK VLNAGEAFVIPR	GLVHFQWNVGQVK VLNAGEAFVIPR	2	99	29	_ (2) -	0	None
O49542, At5g65730 Endoxyloglucan transferase-like protein	SRFPVPPPECR	SRFPVPPPECR	2	99	28	S (3) 30	1 (7–-26*)	None
Q9FLG1, At5g64570 *β*-xylosidase	HYTAYDVDNWK LPMTWYPQSYVEK	HYTAYDVDNWK LPMTWYPQSYVEK	2	99	32	M (4) 102	1 (13–35*)	None
Q9FNQ2, At5g61130 Emb|CAB62612.1 (Putative glycosyl hydrolase family 17 protein)	QSJFNPDNVR SHJNYAVNSFFQK	QS**C**FNPDNVR SH**C**NYAVNSFFQK	2	99	21	S (3) 19	0	None
Q9LU05, At5g44610 Genomic DNA, chromosome 5, TAC clone: K15C23 Hypothetical protein	EGETKPEEIIATGEK KEEAKPVEVPVLAAAEEK	EGETKPEEIIATGEK KEEAKPVEVPVLAAAEEK	2	99	34	_ (4) -	0	None
O81777, At4g31700 Ribosomal protein S6	KGENDLPGLTDTEKPR	KGENDLPGLTDTEKPR	1	99	18	_ (1) -	0	None
CAA61411.1, At4g11650 Osmotin	ILCTADINGQCPNVLR	ILCTADINGQCPNVLR	1	99	31	S (1) 22	1 (5–27*)	None
H84681, At2g28190 Probable copper / zinc superoxide dismutase (imported)	AFVVHELKDDLGK	AFVVHELKDDLGK	1	99	32	C (1) 61	0	None
I39698, At5g20230 Blue copper-binding protein, 20 k	EKPISHMTVPPVK	EKPISHMTVPPVK	1	99	17	S (1) 22	3 (4–21*, 121–143, 173–195)	174, 165
NP_175427.1, At1g50050 Pathogenesis-related protein, putative	HYTQVVWSNSVK	HYTQVVWSNSVK	1	99	23	S (1) 23	0	None
NP_188574.2, At3g19430 Late embryogenesis abundant (LEA) protein-related	TLVAQGFWPYGK	TLVAQGFWPYGK	1	99	26	_ (2) –	0	None
NP_680327.1, At5g35045 Mutator-related transposase	SPCIHAIAAAEHMGVSR	SPCIHAIAAAEHMGVSR	1	99	15	_ (3) –	0	None
O80517, At2g44790 Uclacyanin II precursor (Blue copper-binding protein II)	TVGINYFICSTPGHCR	TVGINYFICSTPGHCR	1	99	27	S (1) 29	2 (10–32*, 184–201)	178, 179

Table 2 continued

Accession/ AGI numbers/ Matched protein	Peptide sequence	Matched sequence	Number of distinct peptides	Best confidence	Score	Target P	TMHMM*	GPI†
Q42342, At5g53560 Cytochrome b5 isoform 1	DATNDFEDVGHSDTAR	DATNDFEDVGHSDTAR	1	99	28	_ (2) –	1 (108–130)	None
Q84WQ6, At5g48490 Hypothetical protein At5g48490 [Fragment]	HADYTCLCGYK	HADYTCLCGYK	1	99	29	S (1) 24	0	None
Q8LDH5, – Endomembrane-associated protein	TEGTSGEKEEIVEETK	TEGTSGEKEEIVEETK	1	99	29	_ (4) –	0	None
Q8LEU7, – Hypothetical protein	EN-PTSPSQPCCTALQHAD-FACLCGYK	EN-PTSPSQPCCTALQHAD-FACLCGYK	1	99	18	S (1) 26	1 (7–29*)	None
Q9FKH4, – Similarity to β-1 Hypothetical protein	DHASFAFNSYYQTYK	DHASFAFNSYYQTYK	1	99	20	S (1) 25	0	None
Q9SKL6, At2g15220 Expressed protein	AGYAPSHWVGPGR	AGYAPSHWVGPGR	1	99	29	S (1) 21	0	None
Q9FF98, At5g23820 Genomic DNA, chromosome 5, P1 clone:MRO11	DGEFTGLLK	DGEFTGLLK	1	98	21	S (1) 24	0	None
NP_175802.1, At1g54000 myrosinase-associated protein, putative	TLVAQGFWPYGK	TLVAQGFWPYGK	1	93	17	S (1) 29	1 (7–29*)	None
BAA19595.1, AtCg00490 Ribulose bisphosphate carboxylase	DLAVEGNEIIR	DLAVEGNEIIR	1	81	22	_ (2) –	0	None
Q8L9M6, – Hypothetical protein	ASGVDPEVALTIPK	ASGVDPEVALTIPK	1	81	18	S (3) 19	0	None
NP_671770.1, At2g03505 glycosyl hydrolase family protein 17	SHCDWAVNTYFQR	SHCDWAVNTYFQR	1	99	15	S (3) 19	0	None
Q9LDB4, At3g08770 Nonspecific lipid-transfer protein 6 precursor (LTP 6)	TIQNALELPSK	TIQNALELPSK	1	99	21	S (1) 19	0	None
O04047, – Putative transcription factor [Fragment]	QILIGANEKENFR	QILIGANEKENFR	1	81	16	_ (2) –	0	None
O22841, At2g43620 Putative endochitinase	YGYCGTTDAYCGTGCR	YGYCGTTDAYCGTGCR	1	99	28	S (2) 28	0	None
Q39131, AT5g15350 Lamin Hypothetical protein	TDYEGCIADHPIR	TDYEGCIADHPIR	1	97	17	S (1) 26	2 (4–26*, 143–165)	None
Q8GYS8, At3g22570 Hypothetical protein	IHTPSFACCSEVYTVGK	IHTPSFACCSEVYTVGK	1	99	20	S (1) 24	0	None

Table 2 continued

Accession/ AGI numbers/ Matched protein	Peptide sequence	Matched sequence	Number of distinct peptides	Best confidence	Score	Target P	TMHMM*	GPI†
Q8LER3, At4g37800 Probable xyloglucan endotransglucosylase/hydrolase protein 7	SRFPVPPPECSAGI	SRFPVPPPECSAGI	1	99	25	S (3) 29	0	None
Q8VXZ7, At3g56310 Putative α-galactosidase	YDNCFNLGIKPIER	YDNCFNLGIKPIER	1	99	22	S (1) 30	1 (12–31*)	None
Q9FMH8, At5g43060 Cysteine protease component of protease-inhibitor complex	ALAHQPISVAIEAGGR	ALAHQPISVAIEAGGR	1	97	21	S (1) 24	0	None
JN0131, At1g02500 Methionine adenosyltransferase	EHVIKPVIPEK	EHVIKPVIPEK	1	81	16	_ (2) –	0	None
Q9M8Y9, At3g04330 Putative trypsin inhibitor	FVDPRPJGESGFWR	FVDPRPJGESGFWR	1	79	16	S (1) 21	0	None
Q9ZUF6, At2g05920 Putative serine protease	SFDDTDOPEIPSK	SFDDTDOPEIPSK	1	55	16	S (3) 21	1(7–26)	None

The proteins with scores below 15 were not included in the list

*Predicted transmembrane helices in the N-terminal region could be a signal peptide. TargetP results are reported in the following order: Location (C = chloroplast, M = Mitochondrian, S = Secretory Pathway, _ = any other location); RC (Reliability class), the lower the RC value, and the safer the prediction; Tplen (predicted length of the presequence). TMHMM results reported in the following order: Number of predicted transmembrane helices; most probable location of transmembrane helices in the sequence (*Predicted transmembrane helices in the N-terminal region could be a signal peptide). GPI results reported in the following order: Potential GPI modification site (sequence position of the omega-site); Potential alternative GPI-modification site

TargetP provides a Reliability Class (RC) value, which is a measure of the size of the difference between the highest and the second highest output scores. Of the 16 identified proteins from *B. napus*, 12 were predicted to be secretory with a RC value of 1. Of the 52 proteins identified from *A. thaliana*, 25 were predicted to be secretory with a RC value of 1 (26 proteins in total had an RC of 1). Six *A. thaliana* proteins were identified as secretory with RC value of 2 (15 proteins in total had an RC of 2).

We identified several proteins in the extracytosolic proteomes of apparently proven function that were not anticipated to reside in the extracellular environment on the basis of previous studies (e.g., CuZnSOD—O81352 in *B. napus* and H84681 in *A. thaliana*). The amino acid sequences also don't show the presence of an identifiable signal peptide that could target them for secretion. However, the possibility of alternate splicing of mRNA, which could be tissue-specific, has to be considered. An alternative transcript could encode a signal peptide, which could account for secretion of the protein. For example, the interleukin-1 receptor antagonist protein exists in two forms, one is intracellular and the other is extracellular. Both forms are derived from the same gene by alternate splicing, resulting in one having a signal peptide for secretion, and the other lacking the signal peptide and so becoming localized intracellularly (Vamvakopoulos et al. 2002). In addition, the inevitable errors in assigning start codons and intron–exon boundaries may also explain the apparent lack of identifiable signal peptides in sequences of the proteins. Furthermore it is increasingly being found that proteins can be located in more than one cellular compartment serving multiple roles (Slabas et al. 2004).

TMHMM (Krogh et al. 2001), a hidden Markov model for predicting transmembrane helices in protein sequences, was used to investigate whether transmembrane domains were predicted in the top hits. The majority of the amino acid sequences that were identified as top hits based on the analysis from both *B. napus* and *A. thaliana* extracellular proteins were predicted to contain no transmembrane domains, which is expected for secreted proteins (Tables 1, 2). A few of the proteins had 1 or 2 hydrophobic domains (e.g., Q9ZUJ8, Q96316 in *B. napus* and Q8LAP0, Q8LEU7 and O80517 in *A. thaliana*). Further analysis of these hydrophobic domains suggested that they are GPI anchor modifications regions (based on big-II plant predictior; Eisenhaber et al. 2003) and/or signal peptides (based on analysis of TargetP results). Glycosylphosphatidylinositol (GPI) anchors are posttranslational modifications, which act to attach proteins to the luminal side of the ER membrane and after vesicular transport to the extracellular leaflet of the plasma membrane. The *B. napus* (Q96316) and *A. thaliana* (O80517 and I39698) proteins that were predicted to contain GPI-anchor modification sites were also predicted to contain secretory signal peptides (RC of 1). Proteins that are GPI-anchored can be released from the plasma membrane to form soluble proteins. The presence of a signal peptide or GPI anchor modification site does not always result in the prediction of a transmembrane domain. Analysis of the big-Π plant predictor results for all of the proteins revealed additional putative GPI-anchored proteins (Q8H794, Q8LE41, and NP_671770.1 in *A. thaliana* and Q9LY37 in *B. napus*), all of which were also predicted to contain signal peptides (RC of 1 or 2, with the exception of NP_671770.1 that had a RC of 3). Five of the *A. thaliana* proteins (Q39131, Q9ZUF6, Q9FLG1, O42342 and I39698) were predicted to contain a hydrophobic region for which the possibility of a transmembrane domain cannot be excluded.

Functional classification of extracellular proteins for the prediction of cellular function

Classification of all proteins according to their function is necessary to get an overview of the functional repertoire of extracellular proteins. We completed a functional classification of the extracellular proteins of both *B. napus* and *A. thaliana* (Table 3) using the Pfam protein family database (Bateman et al. 2004; http://www.sanger.ac.uk/Software/Pfam/). Pfam is a comprehensive collection of protein domains and families, with a range of well-established uses including genome annotation. Pfam families match 75% of protein sequences in Swiss-Prot and TrEMBL. The

Table 3 Functional classification of extracellular root proteins from *B. napus* and *A. thaliana* for the prediction of cellular functions using the Pfam protein family database

Protein family	*B. napus*	*A. thaliana*	Putative function	Reference
Glycoside hydrolases	Q06209, Q8VXZ7	O81862, 023248 AAM63339.1 Q8LF99, Q9FLG1 Q9FNQ2, O22841 O49542, NP_671770.1	Signaling in biotic and abiotic stress conditions	Henrissat et al. (2001), Lee et al. (2003)
Pathogenesis related proteins	–	P33154, P28493, CAA61411.1, NP_175427.1	Induced by biotic and abiotic stresses	Kasprzewska (2003)
Plastocyanin-like domain	Q9ZV18, Q96316	I39698, O80517, Q39131	Aluminum resistance in yeast and *A. thaliana*	Ezaki et al. (2001)
Ubiquitin family	P59263	NP_567291.1	Specific degradation of cell-cycle-regulating proteins of transcription factors, regulation of developmental stress responses	Ingvardsen and Veierskov (2001)
Putative inhibitor/seed storage/LTP family	Q9M8Y9, Q9ZV18	Q84WQ6, Q8LEU7 Q8GYS8, IJXCA, Q8L9M6, Q9ZV18, Q9LDB4	Defense against pathogens, regulation of endogenous storage proteinases during seed dormancy and reserve protein mobilization	Clauss and Mitchell-Olds (2004), Mandal et al. (2002)
Copper/zinc superoxide dismutase (SODC)	O81352; Q9FK60	H84681	Resistance to oxidative stress	Alscher et al. (2002)
Transcription factors	–	Q8GX17 (AP2 domain), O04047	Plant development and hormone-dependent gene expression	Okamuro (1997), Kizis et al. (2001)
Jacalin lectin family proteins	–	Q9ZU19, NP_175617.2, NP_175802.1	Resistance to salt stress, drought and protection against endogenous oxidative damage	Thangstad et al. (2004)
Subtilisin-like protease	–	O65351	Programmed cell death or nutrient scavenging during biotic or abiotic stress	Hamilton et al. (2003)
Class III peroxidase gene family	P24102	–	H_2O_2 detoxification, auxin catabolism, lignin biosynthesis and stress response	Tognolli et al. (2002)
Late embryogenesis abundant proteins (LEA proteins) and supins	–	NP_188574.2	Induced by cold, osmotic stress or exogenous abscisic acid	Wise and Tunnacliffe (2004)
Calmodulin 3 (EF hands)		AAA32765.1	Intracellular Ca2+ receptor involved in transducing a variety of extracellular signals	Yang and Poovaiah (2003)
Plant basic secretory protein	Q9ZUJ8		Defense against biotic stress	Kuwabara et al. (1999)

extracellular proteins we identified were categorized into different families including glycoside hydrolases, pathogenesis-related proteins, trypsin and protease inhibitor, plastocyanin-like domain proteins, copper zinc superoxide dismutases, ubiquitin, protease inhibitor/seed storage/lipid transfer proteins, transcription factors, class III peroxidase gene, and plant basic secretory proteins, BSP (Table 3).

Construction of a web-accessible protein database

In order to facilitate future analysis of the extracellular proteome from other plant species, an online database (http://eppdb.biology.ualberta.ca) containing all of the *B. napus* identified proteins was established (Wang et al. 2004). The database provides an overview of the project, 2-D maps showing the protein locations, and descriptions of the identified proteins (which are further cross-referenced to SWISSPROT/TrEMBL). Protein tables page containing the data from Table 1 giving information about the pI/MW, peptide sequence data obtained after LC-MS/MS and de novo sequencing of all the proteins with the accession number of the top hit is included in the database. We have also included all the protocols used during the course of this project in this database (Wang et al. 2004). The database contains protein entries for all *B. napus* extracellular root proteins identified to date. The format of protein entries is similar to that in SWISS-PROT/TrEMBL and SWISS-2DPAGE. This database will be further expanded with data from *Arabidopsis*.

Advantages and disadvantages of different experimental tools for plant proteome analysis

In this study we obtained the identity of extracytosolic root proteins from both sequenced (*A. thaliana*) and unsequenced (*B. napus*) genomes, using two different mass spectrometric approaches. For most of the proteins identified from *B. napus*, peptide sequences from MS-MS spectra were determined using de novo sequencing algorithms after filtering out low-quality spectra. As manual sequencing and validation of database results is time-consuming and not feasible for the analysis of a large number of proteins, we were successful in obtaining the identity of 16 proteins from a total of 50–60 proteins in the *B. napus* extracellular proteome. In contrast, MudPIT analysis of extracellular proteins from *A. thaliana* resulted in the identification of 52 (from a total of 55–65 proteins).

One of the reasons for the lower number of proteins identified from *B. napus* compared to *A. thaliana* could be the inherent technical limitations of 2-D gel electrophoresis. These include the limited ability to fractionate specific classes of proteins, including hydrophobic proteins and glycoproteins, or the limited dynamic range of the technique, which makes it difficult to visualize low abundance proteins (Rose et al. 2004). The inability to automate several steps in 2-D gel electrophoresis limits throughput and results in greater experimental variability (Rose et al. 2004). Given the limitations, an alternative, 'gel-less' approach (MudPIT) provided an excellent means to identify proteins from *A. thaliana*. However, the connectivity between peptides and proteins is lost when complex protein samples are digested in MudPIT analysis. This loss of connectivity does not exist in identification of proteins extracted from 2D gels based on the digestion of purified protein (Nesvizhskii and Aebersold 2004). Furthermore, in spite of the MudPIT technology being more sensitive, it could not be used for analysis of *B. napus* extracellular proteins in the absence of a sequenced genome. Protein analysis of organisms with incompletely sequenced genomes is a challenging task. Even when using well annotated databases, sequence specific fragmentation or side chain fragmentation of peptides may result in false positive identifications. In this context, de novo sequencing of peptides proved to be a promising tool for the analysis of data of *B. napus*.

In summary, experimental evidence for the existence of specific proteins in root exudates of *A. thaliana* and *B. napus* can serve as a platform for future work to investigate the role of extracellular proteins in the development and stress resistance of plants. Engineering the rhizosphere to elicit benefits such as suppression of soil-borne crop diseases, phytoremediation, optimization of

nutrient supply, production of plant growth promoting substances, and improvement of plant–soil–water relations offers a new approach to land management. Practical manipulation and engineering of the rhizosphere can only be facilitated by the development of a better understanding of biological interactions in the root environment.

Acknowledgements This work was funded by an Alberta Agricultural Research Institute (AARI) grant to GJT and RMW, Natural Sciences and Engineering Research Council of Canada (NSERC) grant to GJT, RG, OZ and AGG and FGAS project to DGM, AGG and GJT. The FGAS project is supported by Genome Prairie, in part through Genome Canada, a not-for-profit corporation which is leading a national strategy on genomics with $375 million in funding from the Government of Canada. The personnel at the Institute for Biomolecular Design (IBD), University of Alberta, Edmonton, AB, Canada and University of Victoria BC Proteomics Centre staffs are gratefully acknowledged for their expertise and assistance with mass spectrometry.

References

Alscher RG, Erturk N, Heath LS (2002) Role of superoxide dismutases (SODs) in controlling oxidative stress in plants. J Exp Bot 53:1331–1341

Bais HP, Loyola VVM, Flores HE, Vivanco JM (2001) Root specific metabolism: the biology and biochemistry of underground organs. In Vitro Cell Dev Biol Plant 37:730–741

Bais HP, Park SW, Weir TL, Callaway RM, Vivanco JM (2004) How plants communicate using the underground information superhighway. Trends Plant Sci 9:26–32

Basu U, Basu A, Taylor GJ (1994) Differential exudation of polypeptides by roots of aluminum-resistant and aluminum-sensitive cultivars of *Triticum aestivum* L. in response to aluminum stress. Plant Physiol 106:151–158

Basu U, Good AG, Aung T, Slaski JJ, Basu A, Briggs KG, Taylor GJ (1999) A 23 kD, aluminum-binding, root exudates polypeptide co-segregates with the aluminum-resistant phenotype in *Triticum aestivum*. Physiol Plant 106:53–61

Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer ELL, Studholme DJ, Yeats C, Eddy SR (2004) The Pfam protein families database. Nucleic Acids Res 32:D138–D141

Borisjuk NV, Borisjuk LG, Logendra S, Petersen F, Gleba Y, Raskin I (1999) Production of recombinant proteins in plant root exudates. Nat Biotechnol 17:466–469

Bradford MM (1976) A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. Anal Biochem 72:248–254

Briat JF, Lebrun M (1999) Plant responses to metal toxicity. C R Acad Sci III 322:43–54

Chang WWP, Huang L, Shen M, Webster C, Burlingame AM, Roberts JKM (2000) Patterns of protein synthesis and tolerance of anoxia in root tips of maize seedlings acclimated to a low-oxygen environment and identification of proteins by mass spectrometry. Plant Physiol 122:295–317

Chuong SDX, Good AG, Taylor GJ, Freeman MC, Moorhead GBG, Muench DG (2004) Large-scale identification of tubulin binding proteins provides insight on subcellular trafficking, metabolic channeling, and signaling in plant cells. Mol Cell Proteomics 3:970–983

Clauss MJ, Mitchell-Olds T (2004) Functional divergence in tandemly duplicated *Arabidopsis thaliana* trypsin inhibitor genes. Genetics 166:1419–1436

Crofts AJ, Leborgne-Castel N, Hillmer S, Robinson DG, Phillipson B, Carlsson LE, Ashford DA, Denecke J (1999) Saturation of the endoplasmic reticulum retention machinery reveals anterograde bulk flow. Plant Cell 11:2233–2247

Drake PMW, Chargelegue DM, Vine ND, Dolleweerd CJV, Obregon P, Ma JKC (2003) Rhizosecretion of a monoclonal antibody protein complex from transgenic tobacco roots. Plant Mol Biol 52:233–241

Dreger M (2003) Proteome analysis at the level of subcellular structures Eur J Biochem 270:589–599

Eisenhaber B, Wildpaner M, Schultz CJ, Borner GHH, Dupree P, Eisenhaber F (2003) Glycosylphosphatidylinositol lipid anchoring of plant proteins. Sensitive prediction from sequence- and genome-wide studies of Arabidopsis and Rice. Plant Physiol 133:1691–1701

Emanuelsson O, Nielsen H, Brunak S, Von Hejne G (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. J Mol Biol 300:1005–1016

Ezaki B, Katsuhara M, Kawamura M, Matsumoto H (2001) Different mechanisms of four aluminum (Al)-resistant transgenes for Al toxicity in *Arabidopsis*. Plant Physiol 127:918–927

Flores HE, Vivanco JM, Loyola-Vargas VM (1999) 'Radicle' biochemistry: the biology of root-specific metabolism. Trends Plant Sci 4:220–226

Gleba D, Borisjuk NV, Borisjuk LG, Kneer R, Poulev A, Skarzhinskaya M, Dushenkov S, Logendra S, Gleba YY, Raskin I (1999) Use of plant roots for phytoremediation and molecular farming. Proc Natl Acad Sci USA 96:5973–5977

Hamilton JMU, Simpson DJ, Hyman SC, Ndimba BK, Antoni R, Slabas AR (2003) Ara12 subtilisin-like protease from *Arabidopsis thaliana*: purification, substrate specificity and tissue localization. Biochem J 370:57–67

Henrissat B, Coutinho M, Davies GJ (2001) A census of carbohydrate-active enzymes in the genome of *Arabidopsis thaliana*. Plant Mol Biol 47:55–72

Ingvardsen C, Veierskov B (2001) Ubiquitin-and proteasome-dependent proteolysis in plants. Physiol Plant 112:451–459

Jasinski M, Stukkens Y, Degand H, Purnell B (2002) A plant plasma membrane ATP binding cassette-type transporter is involved in antifungal terpenoid secretion. Plant Cell 13:1095–1107

Kasprzewska A (2003) Plant Chitinases-regulation and function. Cell Mol Biol Lett 8:809–824

Kizis D, Lumbreras V, Pages M (2001) Role of AP2/EREBP transcription factors in gene regulation during abiotic stress. FEBS Lett 498:187–189

Krogh A, Larsson B, Heijne GV, Sonnhammer ELL (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol 305:567–580

Kuwabara C, Arakawa K, Yoshida S (1999) Abscisic acid-induced secretory proteins in suspension-cultured cells of winter wheat. Plant Cell Physiol 40:184–191

Lee HY, Bahn SC, Kang YM, Lee KH, Kim HJ, Noh EK, Palta JP, Shin JS, Ryu SB (2003) Secretory low molecular weight phospholipase A2 plays important roles in cell elongation and shoot gravitropism in *Arabidopsis*. Plant Cell 15:1990–2002

Mandal S, Kundu P, Roy B, Mandal RK (2002) Precursor of the inactive 2S seed storage protein from the Indian mustard *Brassica juncea* is a novel trypsin inhibitor. J Biol Chem 277:37161–37168

Medzihradszky KF, Burlingame AL (1994) The advantages and versatility of a high-energy collision-induced dissociation-based strategy for the sequence and structural determination of proteins. Methods: A comparison to Methods in Enzymology 6:284–303

Nardi S, Concheri G, Pizzeghello D, Sturaro A, Rella R, Parvoli F (2000) Soil organic matter mobilization by root exudates. Chemosphere 5:653–658

Nesvizhskii AI, Aebersold R (2004) Analysis, statistical validation and dissemination of large-scale proteomics datasets generated by tandem MS. DDT 9:173–181

Okamuro JK, Caster B, Villarroel R, Montagu MV Jofuku KD (1997) The AP2 domain of *APETALA2* defines a large new family of DNA binding proteins in *Arabidopsis*. Proc Natl Acad Sci USA 94:7076–7081

Pandey A, Mann M (2000) Proteomics to study genes and genomes. Nature 405:837–846

Park OK (2004) Proteomic studies in plants. J Biochem Mol Biol 37:133–138

Peng J, Elias JE, Thoreen CC, Licklider LJ, Gygi SP (2003) Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome. J Proteome Res 2:43–50

Raghothama KG (1999) Phosphate acquisition. Annu Rev Plant Physiol Plant Mol Biol 50:665–693

Richards KD, Schott EJ, Sharma YK, Davis KR, Gardner R (1998) Aluminum induces oxidative stress genes in *Arabidopsis thaliana*. Plant Physiol 116:409–418

Rose JKC, Bashir S, Giovannoni JJ, Jahn MM, Saravanan RS (2004) Tackling the plant proteome: practical approaches, hurdles and experimental tools. Plant J 39:715–733

Shepherd T, Davies HV (1993) Carbon loss from the roots of forage rape (*Brassica napus* L.) seedlings following pulse-labeling with CO_2. Ann Bot 72:155–163

Shevchenko A, Wilm M, Vorm O, Mann M (1996) Mass spectrometric sequencing of proteins from silver-stained polyacrylamide gels. Anal Chem 68:850–858

Slabas AR, Ndimba B, Simon WJ, Chivasa S (2004) Proteomic analysis of the *Arabidopsis* cell wall reveals unexpected proteins with new cellular locations. Biochem Soc Trans 32: part 3

Thangstad OP, Gilde B, Chadchawan S, Seem M, Husebye H, Bradley D, Bones AM (2004) Cell specific, cross-species expression of myrosinases in *Brassica napus*, *Arabidopsis thaliana* and *Nicotiana tabacum*. Plant Mol Biol 54:597–611

Tognolli M, Penel C, Greppin H, Simon P (2002) Analysis and expression of the class III peroxidase large gene family in *Arabidopsis thaliana*. Gene 288:129–138

Tomscha JL, Trull MC, Deikman J, Lynch JP, Guiltinan MJ (2004) Phosphatase under-producer mutants have altered phosphorus relations. Plant Physiol 135:334–345

Vamvakopoulos JE, Taylor CJ, Morris-Stiff GJ, Green C, Metcalfe S (2002) The interleukin-1 receptor antagonist gene: a single-copy variant of the intron 2 variable number tandem repeat (VNTR) polymorphism. Eur J Immunogenetics 29:337–340

Vitale A, Denecke J (1999) The endoplasmic reticulum-gateway of the secretory pathway. Plant Cell 11:615–628

Walker TS, Bais HP, Grotewold E, Vivanco JM (2003) Root exudation and rhizosphere biology. Plant Physiol 132:44–51

Wang Y, Zaiane O, Goebel R, Southron JL, Basu U, Whittal RM, Stephens JL, Taylor GJ (2004) Developing a database for proteomic analysis of extracytosolic plant proteins. Second International Workshop on Biological Data Management (BIDM'2004) in conjunction with the 15th Int. Conf. on Database and Expert Systems Applications DEXA2004, Zaragoza, Spain, August 30–September 3, 2004

Whitelegge JP (2003) Plant proteomics: blasting out of a MudPIT. Proc Natl Acad Sci 99:11564–11566

Wise MJ, Tunnacliffe A (2004) POPP the question: what do LEA proteins do?. Trends Plant Sci 9:13–17

Yang T, Poovaiah BW (2003) Calcium/calmodulin-mediated signal network in plants. Trends Plant Sci 10:505–512