Available online at www.sciencedirect.com

SCIENCE DIRECT®

Phytochemistry 65 (2004) 1733–1744

PHYTOCHEMISTRY

www.elsevier.com/locate/phytochem

High-throughput peptide mass fingerprinting of soybean seed proteins: automated workflow and utility of UniGene expressed sequence tag databases for protein identification

Brian P. Mooney, Jay J. Thelen *

Department of Biochemistry and Proteomics Center, University of Missouri-Columbia, 125 Chemistry, Columbia, MO 65211, USA

Received 9 January 2004; received in revised form 1 April 2004
Available online 7 May 2004

Abstract

Identification of anonymous proteins from two-dimensional (2-D) gels by peptide mass fingerprinting is one area of proteomics that can greatly benefit from a simple, automated workflow to minimize sample contamination and facilitate high-throughput sample processing. In this investigation we outline a workflow employing robotic automation at each step subsequent to 2-D gel electrophoresis. As proof-of-concept, 96 protein spots from a 2-D gel were analyzed using this approach. Whole protein (1 mg) from mature, dry soybean (*Glycine max* [L.] Merr.) cv. Jefferson seed was resolved by high resolution 2-D gel electrophoresis. Approximately 150 proteins were observed after staining with Coomassie Blue. The rather low number of detected proteins was due to the fact that the dynamic range of protein expression was greater than 100-fold. The most abundant proteins were seed storage proteins which in total represented over 60% of soybean seed protein. Using peptide mass fingerprinting 44 protein spots were identified. Identification of soybean proteins was greatly aided by the use of annotated, contiguous Expressed Sequence Tag (EST) databases which are available for public access (UniGene, ftp.ncbi.nih.gov/repository/UniGene/). Searches were orders of magnitude faster when compared to searches of unannotated EST databases and resulted in a higher frequency of valid, high-scoring matches. Some abundant, non seed storage proteins identified in this investigation include an isoelectric series of sucrose binding proteins, alcohol dehydrogenase and seed maturation proteins. This survey of anonymous seed proteins will serve as the basis for future comparative analysis of seed-filling in soybean as well as comparisons with other soybean varieties.

Keywords: Soybean seed; Proteomics; 2-D electrophoresis; MALDI-TOF mass spectrometry; Peptide mass fingerprinting; Robotic automation

1. Introduction

Current two-dimensional (2-D) gel electrophoresis in combination with advanced mass spectrometry has revolutionized the large-scale profiling of proteins, also referred to as proteomics (Wilkins et al., 1995). Protein spots isolated from 2-D polyacrylamide gels and digested with proteases generate reproducible peptide fragments whose molecular mass can be accurately and rapidly determined using Matrix-Assisted Laser Desorption Ionization Time-of-Flight (MALDI-TOF) mass spectrometry. These experimentally determined peptide masses can be utilized to search "in silico" digested protein or nucleic acid databases for statistical matches (Yates et al., 1993; Cottrell, 1994; Wilkins et al., 1995). This systematic approach for the identification of unknown proteins based upon peptide masses is referred to as peptide mass fingerprinting (PMF; Cottrell, 1994). Although PMF is a facile approach well-suited for au-

Abbreviations: ADH, alcohol dehydrogenase; 2-D, two-dimensional; CHAPS, 3-[(3-cholamidopropyl) dimethylammonio]-1-propanesulfonate; CBB, Coomassie Brilliant Blue; DTT, dithiothreitol; EST, expressed sequence tag; FTP, file transfer protocol; IPG, immobilized pH gradient; MALDI-TOF, matrix-assisted laser desorption ionization time-of-flight; MS, mass spectrometry; NCBI, National Center for Biotechnology Information; PMF, peptide mass fingerprinting; SBP, sucrose binding protein; Vh, volt-hours.

* Corresponding author. Tel.: +1-573-884-7374; fax: +1-537-882-2754.

E-mail address: thelenj@missouri.edu (J.J. Thelen).

0031-9422/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.phytochem.2004.04.011

tomation it is strongly dependent upon high quality protein databases for making accurate assignments (Cottrell, 1994; Wilkins et al., 1995; Mann et al., 2001). Since extensive genome sequence information is available for some mammals and microbes, PMF is the method of choice for high-throughput identification of proteins from these organisms. Indeed, peptides can be matched to raw genome data without assumptions about reading frames or coding regions (Pandey and Mann, 2000; Mann et al., 2001). Thus peptide mass fingerprinting approaches can also help define gene structure in the post-genome era (Mann et al., 2001).

Previous reports have demonstrated that in the absence of genomic sequence data, Expressed Sequence Tag (EST) databases can be utilized in PMF investigations to identify proteins (Mann et al., 2001; Porubleva et al., 2001; Watson et al., 2003). Although sequencing of the soybean genome is in its infancy, substantial EST sequence information is currently available. Public soybean EST projects have generated 344,524 EST nucleotide sequences as of December 2003 (ncbi.nlm.nih.gov/dbEST_summary.html). However, due to their poor sequence quality and short length (474 nucleotide average), unannotated ESTs are generally unsuitable for proteome investigations. For these reasons cluster analysis of ESTs to assemble databases of contiguous cDNA sequences is a potentially useful resource for proteome investigations with organisms lacking sequenced and annotated genomes. These datasets, also referred to as UniGene databases, are available in FASTA format for public access as a resource of the National Center for Biotechnology Information (NCBI, Pontius et al., 2003).

Soybeans supply a major portion of the world's demand for vegetable oil and protein. Salt-soluble globulins, the predominant seed storage proteins, are classified as either 7S or 11S based upon their sedimentation coefficients on sucrose gradients (Thanh and Shibasaki, 1976a). These proteins account for approximately 50-70% of the total protein within the soybean meal (Thanh and Shibasaki, 1976a,b). Sodium dodecyl sulfate-polyacrylamide gel electrophoresis of soybean seed proteins extracted with six different solvent systems have resulted in the resolution of 292 polypeptides (Hu and Esen, 1981). Heterogeneity was further verified by 2-D polyacrylamide gel electrophoresis (Hu and Esen, 1982) in which hundreds of polypeptide spots were detected. The major polypeptides were tentatively identified as components of 7S and 11S globulins, however, the vast majority of these soybean seed proteins remain to be identified.

The following proteomic survey of soybean seed proteins extends previous research by assigning identity to some of the previously observed proteins on 2-D gels. We also discuss a simplified, automated workflow for PMF and demonstrate the utility of UniGene databases in proteomic investigations. Furthermore, a method for protein isolation from recalcitrant tissues such as soybean seed is described in detail. This procedure was developed to enable milligram quantities of whole protein to be resolved using the current immobilized pH gradient (IPG) strip method for isoelectric focusing.

2. Experimental

2.1. Isolation of soybean seed protein for two-dimensional electrophoresis

Total protein from soybean seed was isolated by a procedure from Hurkman and Tanaka (1986) with the following modifications. Dry seed (1 g) was pulverized to a fine powder in a mortar and pestle in the presence of liquid nitrogen. Powder was resuspended directly in the mortar with 15 mL of homogenization media (50% phenol, 0.45 M sucrose, 5 mM EDTA, 0.2% (v/v) 2-mercaptoethanol, 50 mM Tris–HCl pH 8.8) with continued homogenization until homogenate reached room temperature. Homogenate was transferred to a phenol-resistant screw cap tube and mixed on a Nutator mixer for 30 min at 4 °C. Homogenate was centrifuged at 5000*g* for 15 min at 4 °C in a swinging bucket rotor. The top phenol phase was removed and added to five volumes of ice cold 0.1 M ammonium acetate in 100% methanol and mixed before placing at −20 °C for a minimum of 1 h. Precipitated protein was collected by centrifugation (10 min at 5000*g*) and supernatant was decanted. The precipitate was thoroughly washed twice in 20 mL of 0.1 M ammonium acetate in 100% methanol followed by two washes with ice-cold 80% acetone and a final wash in ice-cold 70% ethanol. Washed protein pellet was either stored as a precipitate at −20 °C or dried and resuspended for immediate isoelectric focusing.

2.2. Isoelectric focusing using immobiline pH gradient strips

Protein pellet (from 1 g of dry seed material) was resuspended in 2 mL of isoelectric focusing resuspension media (8 M urea, 2 M thiourea, 2% 3-[(3-cholamidopropyl) dimethylammonio]-1-propanesulfonate (CHAPS), 2% Triton X-100, 50 mM DTT, 2 mM tributyl phosphine, 0.5% carrier ampholytes) by pipetting followed by vortexing (1 h at 10% maximum). Insoluble matter was sedimented by centrifugation (20 min at 14,000*g*) and supernatant was removed and placed into a separate tube. Protein concentration was determined using the dye-binding assay (Bio-Rad, Hercules, CA) using only 0.5–2 μL of sample to minimize interference from high concentrations of detergent and reductant. Protein quantitation was performed in triplicate, and quantitated against a standard curve of chicken gamma-globulin standard. Exactly 1.0 mg of protein was added to a separate tube and volume was brought up to 0.25 mL with IEF resuspension media and mixed before pipetting into a 13 cm ceramic IPG rehydration tray (Amersham Biosciences, Piscataway, NJ). Immobiline pH gradient (IPG, 13 cm) strips were carefully placed onto resuspended protein samples which were then overlayed with 1.5 mL of mineral oil. Active rehydration (12 h at 50 V) was performed and this was directly linked to a three step focusing protocol on an IPGphor IEF unit (Amersham Biosciences) 500 V for 750 Vh, 1000 V for 2000 Vh, and 8000 V for 24 kVh.

2.3. SDS–PAGE for 2-D electrophoresis

Following isoelectric focusing, IPG strips were removed from the focusing tray and blotted on kimwipes to remove mineral oil. The strips were then incubated in equilibration buffer (50 mM Tris pH 6.8, 6 M urea, 30% (v/v) glycerol, 5% (w/v) SDS) supplemented with 2% (w/v) DTT for 15 min with gentle agitation, followed by incubation in buffer supplemented with 2.5% (w/v) iodoacetamide for 15 min with gentle agitation. IPG strips were then rinsed with SDS–PAGE running buffer and placed onto 11–17% linear acrylamide gradient gels. Strips were then overlayed with agarose solution (60 mM Tris–HCl pH 6.8, 60 mM SDS, 0.5% (w/v) agarose, 0.01% (w/v) bromophenol blue). Second dimension SDS–PAGE was conducted in Hoeffer SE600 units for 4 h at 30 mA constant current per gel. Following SDS–PAGE gels were washed with deionized water three times 15 min each and stained for at least 12 h with colloidal Coomassie (20% (v/v) ethanol, 1.6% (v/v) phosphoric acid, 8% (w/v) ammonium sulfate, 0.08% (w/v) Coomassie Brilliant Blue G-250). Each soybean seed preparation was resolved by 2-D electrophoresis in at least three independent experiments.

2.4. Image analysis, spot excision and tryptic digestion of proteins

Electronic images of two-dimensional gels were analyzed using Phoretix 2D-Advanced software (Nonlinear Dynamics, Newcastle, UK). Spot detection, background subtraction and spot quantitation were performed on 16-bit TIFF images acquired with a scanning densitometer. Background subtraction on detected spots was performed using the mode-of-non-spot function. Protein spots were excised and arrayed into 96-well MultiScreen^{TM} model R5, 5 μM hydrophilic PTFE membrane, glass-filled polypropylene plates (Millipore, Bedford, MA) using 1.4 mm diameter pins on the GelPix robotic spot excision station (Genetix Ltd., UK).

After arraying of protein spots, 96-well plates were transferred to a Multiprobe II EX liquid handling station (Packard) for subsequent destaining, tryptic digestion and peptide extraction. Gel plugs were destained in 200 μL of 50% (v/v) acetonitrile, 50 mM ammonium bicarbonate and incubated at room temperature (25 °C) for 30 min. Destain solution was evacuated from the bottom of the filter plates using a vacuum manifold configured specifically for the Multiprobe robot (Millipore, Bedford, MA). Destaining was repeated two times or until Coomassie stain was removed from the gel plugs. Gel plugs were dehydrated in 100% acetonitrile for 5 min at room temperature. Acetonitrile was evacuated from the plates using a vacuum manifold and underside of plates were gently blotted to filter paper to remove residual acetonitrile. Sequencing grade trypsin (20 μg, Promega, Madison, WI) was thoroughly resuspended in 5 mL of 50 mM ammonium bicarbonate and 50 μL aliquoted into the 96-well plate. Adhesive tape was placed over the wells, a 96-well V-bottom sample collection plate was placed underneath the MultiScreen plate (to collect any liquid) and the samples were incubated at 37 °C for 16 h to allow for complete digestion. Peptides were extracted from the gel plugs with 50 μL of 60% (v/v) acetonitrile, 0.3% (v/v) trifluoroacetic acid and agitation in a microplate shaker (140 rpm) for 15 min. Tryptic peptides were collected into the V-well collection plate using a vacuum evacuation manifold. Extracted peptides were concentrated to 5–15 μL by centrifugal vacuum evaporation.

2.5. Mass determination of tryptic peptides

Tryptic peptide samples (0.5 μL) were applied to a 96 × 2 Teflon® MALDI plate using a Symbiot I liquid handling station (Applied Biosystems, Inc., Foster City, CA). The samples were mixed on-target with an equal volume of the matrix solution, 10-mg/mL α-cyano-4-hydroxycinnamic acid (Sigma–Aldrich Fluka, St. Louis, MO) prepared in 60% (v/v) acetonitrile, 0.3% (v/v) trifluoroacetic acid. Analyses of trypsin-digested protein samples were carried out on a Voyager-DE Pro Matrix Assisted Laser Desorption Ionization-Time of Flight mass spectrometer (MALDI-TOF MS, Applied Biosystems). The MALDI-TOF MS was operated in the positive ion delayed extraction reflector mode for highest resolution and mass accuracy. Peptides were ionized/desorbed with a 337-nm laser and spectra were acquired at 20 kV accelerating potential with optimized parameters. The close external calibration method employing a mixture of standard peptides (Applied Biosystems) provided mass accuracy of 25–50 ppm across the mass range of 700–4500 Da.

Peptide spectra were automatically processed for baseline correction, noise removal, peak deisotoping and threshold adjustment (2% base peak intensity) prior to submission to a local copy of version 3.2.1 of the MS Fit program of Protein Prospector (http://prospector.ucsf.edu). Search criteria required the match of at least four peptides with a mass error of less than 100 ppm for a tentative protein assignment. Each protein assignment from Protein Prospector was confirmed using the 'IntelliCal' search function of Proteomics Solution 1 (Applied Biosystems) which recalibrates spectra according to top scoring peptide hits for a further round of database searching. Assignments from UniGene contigs were subsequently searched against the NCBI non-redundant database using the BLASTP search algorithm to determine similarity matches.

3. Results and discussion

3.1. Two-dimensional electrophoresis of whole seed protein from soybean

Development of a robust protein isolation protocol that is compatible with isoelectric focusing was integral to developing reproducible 2-D gel proteome reference maps of soybean seed protein. Preliminary experiments revealed that protein extracted from soybean seed, in native or denaturing conditions, followed by acetone precipitation of protein prior to a final resuspension in IEF extraction media resulted in a highly viscous protein sample. Analysis of this protein fraction by 2-D electrophoresis (0.5–1 mg protein loading) resulted in substantial horizontal streaking which suggested contamination with polysaccharides, nucleic acids or other non-protein macromolecules (data not shown). Insertion of a phenol extraction step immediately after seed crushing effectively partitioned the protein sample from these interfering contaminants as observed by reduced viscosity in the final sample and minimal horizontal streaking on 2-D gels at greater than 0.5 mg protein loadings.

Comparison of seed protein (1 mg) from two soybean varieties, Jefferson and Williams 82, resolved by 2-D electrophoresis revealed strong similarities in the overall distribution and polypeptide expression pattern (Fig. 1). These 2-D gels for whole soybean seed were also similar to previous gels which used petroleum ether defatted soybean meal as the starting material (Hu and Esen, 1982). The high level of similarity indicated reproducibility in the protein extraction and electrophoretic separation methods. Despite the overall similarity in protein expression between the two varieties in Fig. 1, at least twenty differences were detected from experimental triplicate analyses. Verification of these differences will also require 2-D electrophoresis in biological triplicate to perform statistical analyses. Although the loading of high quantities of protein was necessary to observe lower abundance proteins this frequently resulted in the coalescing of abundant isoelectric protein series (Fig. 1, top panel). Analysis of a lower protein loading (0.1 mg) of Jefferson soybean demonstrated multiple isoelectric forms for three different classes of proteins (Fig. 1, right panels). The tight clustering of these series is suggestive of post-translational modification rather than genetic isoforms.

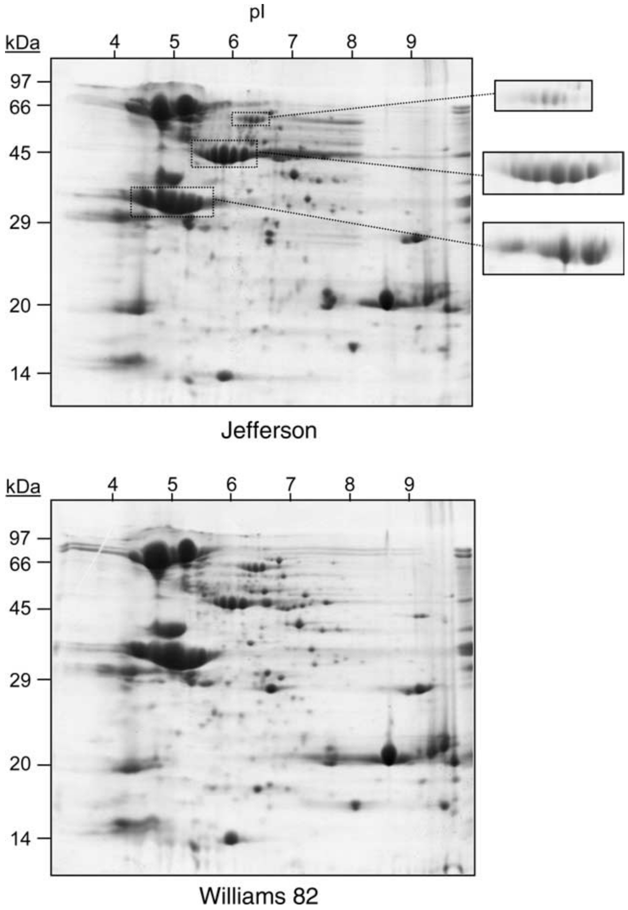


Fig. 1. Comparison of two-dimensional gel proteome maps of soybean (*Glycine max*) mature seed varieties. Total protein (1 mg each) from Jefferson (top panel) and Williams 82 (bottom) varieties were resolved by high-resolution two-dimensional gel electrophoresis and stained with Coomassie Blue. Seed protein was resolved on pH 3–10 linear IPG strips in the first dimension and 11–17% linear polyacrylamide gels in the second dimension. pH range and molecular masses are noted. A lower protein loading (100 μg) of Jefferson seed variety revealed distinct isoelectric "trains" for at least two sets of proteins (right panels).

Detailed analysis of the CBB-stained 2-D electrophoretic map of Jefferson soybean seed showed 128 distinct, intense protein spots (Fig. 2). Of these spots, 96 were quantitated using Phoretix 2-D Advanced software and subjected to automated peptide mass fingerprinting. Semi-quantitative analysis of these spots was performed by resolving protein standards (at five concentrations) next to the IPG strips on the second dimension gel. Each protein on this 2-D map except for the four most abundant spots (4, 8, 28, 45) were within the linear quantitative range of these standards. Each spot was normalized

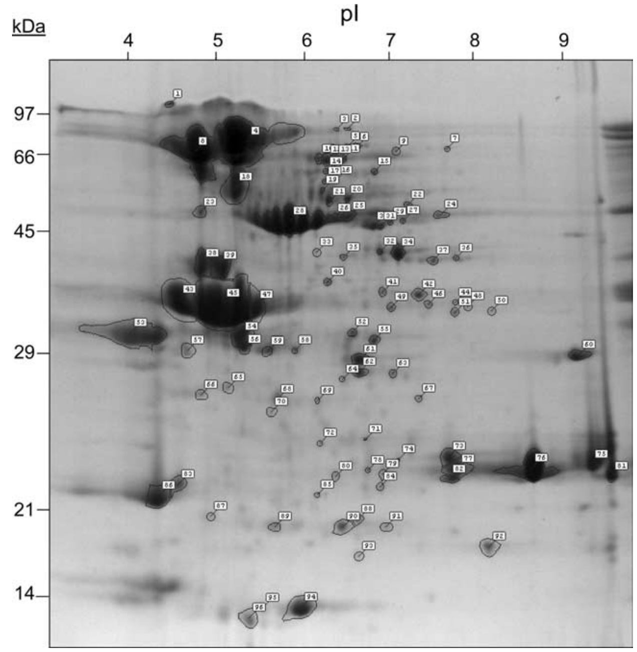


Fig. 2. Two-dimensional spot-excision map of mature soybean seed, Jefferson variety. Total protein (1 mg) was resolved on pH 3–10 linear IPG strip followed by 11–17% linear SDS–PAGE. Mass ranges and pH are noted. The Coomassie-stained gel was imaged by densitometry scanning and proteins were detected and quantitated using Phoretix 2D Advanced software. Exactly 96 protein spots were chosen for spot-excision and subsequent peptide mass fingerprint identification. Protein spots were manually chosen for analysis based upon intensity, circularity and resolution.

to spot 76 which was quantified to be 60 μg. Dynamic range of protein expression and accumulation was greater than 1500-fold based upon volume comparisons of the highest and lowest intensity spots. Although this broad dynamic range is the result of a few seed storage proteins it is problematic for seed proteome investigations because the preponderance of these proteins masks the detection of other proteins. Since the amount of protein that can be loaded onto an IPG strip is limited to 1–2 mg, based on our experience, in-depth soybean seed proteome investigations using 2-D electrophoresis will likely require pre-fractionation steps or narrow range IPG strips for isoelectric focusing. Narrow pH range IPG strips allow for greater protein loads to increase the number of detectable spots. Although fluorescent dye detection methods are reportedly greater than 2-fold more sensitive than CBB (Steinberg et al., 2000), in our hands, only 10–15% more spots are typically observed (data not shown). Alternatively, subtraction of abundant proteins using affinity chromatography could "unmask" lower abundance proteins and this approach appears to be promising for the removal of albumin and IgG animal sera (Pieper et al., 2003). However, it is questionable whether this approach could be utilized for seed proteome investigations as storage proteins are assembled into complex, membrane-bound protein bodies (Herman and Larkins, 1999).

3.2. *Workflow for automated peptide mass fingerprinting from two-dimensional gels*

As the frequency and scope of proteome investigations increase, robotic laboratory automation has become critical to maximize efficiency and reproducibility while minimizing handling errors during arraying and sample tracking. The overall workflow for automated PMF from 2D gels is shown in Fig. 3. Each step subsequent to 2D electrophoretic separation is automated (Fig. 3). Robotic automation for protein identification by PMF can be divided into at least three categories: spot excision and arraying, liquid handling for in-gel digestion and MALDI plate spotting. Although image capture and analysis could be considered a fourth application for robotic instrumentation, the pixel depth, resolution and sensitivity of images acquired from onboard CCD cameras or flatbed scanners are frequently inferior to offline imaging devices such as high-resolution image densitometers and laser scanners. Although unsuitable for quantitative investigations, the robotic image capture devices are very useful for verifying excision accuracy and also aligning (triangulating) offline images to current images for spot-excision.

Since each of the aforementioned applications requires specific sampling and handling capabilities most robotic instrumentation is only capable of performing one or two of these functions. However, recently released integrated or 'hybrid' instruments are capable of performing each of these three applications in one contained unit. Multi-function robots generally conserve resources and lab space, however, the sample throughput with these instruments is typically low. Since there is a wide variety of robotic automation for proteomics applications and each instrument has different capabilities as well as workflows a summary is provided in Table 1. The intent of this table is to provide an overview of the instrumentation that is currently available and does not comprehensively detail the features or capabilities of each unit. Besides imaging capabilities and sample throughput, MALDI plate configurability and the ability to integrate into a seamless workflow are two important considerations, particularly when implementing multiple robots from different vendors.

In this investigation, a total of 96 spots (shown in Fig. 2) were subjected to automated PMF analysis. Protein spots were excised and arrayed into a polypropylene PVDF filter-bottom Multiscreen plate using a GelPix spot-picking robot. The GelPix robot excises and arrays polyacrylamide spots at a rate exceeding 400 spots/h using a novel 8-pin cutting head. With a capacity of fifteen 96-well plates, as many as 1440 gel plugs can be arrayed in a single, uninterrupted run. This capacity

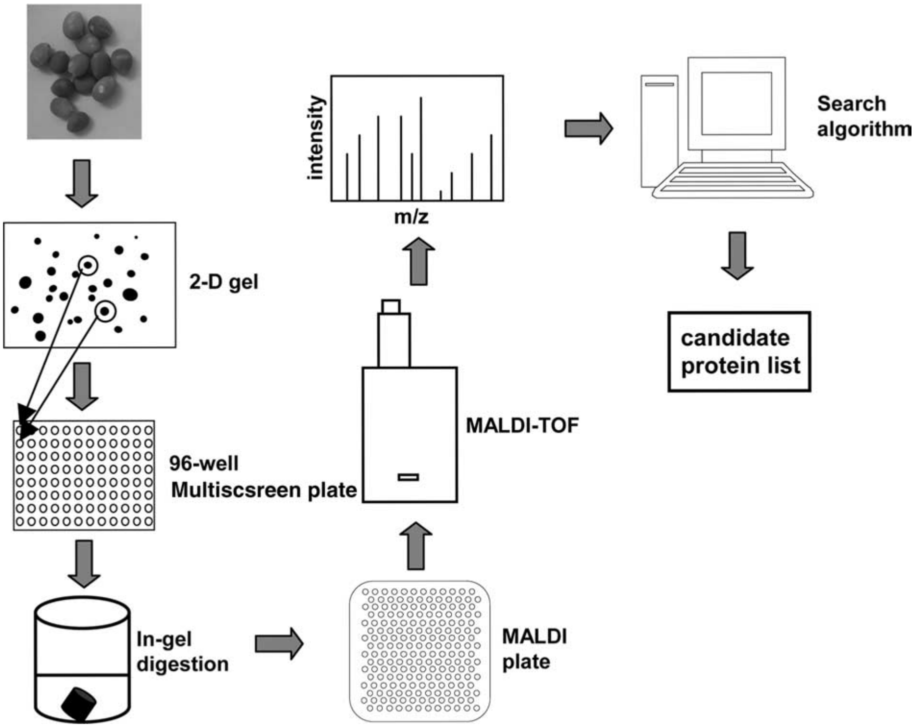


Fig. 3. Diagram of workflow for automated peptide mass fingerprinting protein identification from two-dimensional gels. Arraying of gel spots into 96-well polypropylene Multiscreen plates (Millipore, Inc.) was performed using a spot-cutting robot (GelPix, Genetix, Ltd.) with an enclosed, HEPA-filtered workspace. Liquid handling for in-gel digestions was performed using a robotic liquid handling station (Multiprobe II, Packard). Resuspension and direct spotting of lyophilized tryptic peptides was performed with a low-volume liquid handling robot (Symbiot I, Applied Biosystems, Inc.). Automated mass spectral acquisitions were acquired with a Voyager DE-PRO MALDI-TOF mass spectrometer workstation and spectra were processed and queued directly into a local copy of Protein Prospector for database querying. Sample tracking of data from gel analysis through database searching was compiled into a Microsoft Access database through the Proteomics Solutions 1 software suite (Applied Biosystems, Inc.).

has thusfar been adequate for our workflow as we have never resolved and detected more than 1300 CBB- or Sypro Ruby-stained spots in a single 24 cm format 2D gel. Excision efficiency and accuracy for all 96 spots excised in this investigation was verified by comparing images pre- and post-excision using an on-board CCD camera.

In-gel digestions were performed using a Multiprobe II liquid handling robot equipped with a vacuum manifold to evacuate solutions from the bottom of the filter-bottom 96-well plate. Although versatile and reliable for volumes greater than 5 μL, the poor throughput and plate capacity of this robot makes it extremely cumbersome for processing multiple 96-well plates. Availability of an add-on gripper arm which is capable of transfering sample plates anywhere on the platform deck may allow for true walk-away automation for multiple plate processing. With our current workflow, liquid handling for multiple plates of in-gel digestions is most easily performed manually using an electronic multichannel pipetter operated within a laminar flow hood, to minimize airborne contamination. Although the Multiprobe can be configured for MALDI plate spotting we encountered difficulty spotting the 192-well MALDI plates from Applied Biosystems. Due to the small plate and target size, the maximum total volume that could be spotted onto the target well of a 192-well plate is 1 μL. In our hands, the dispensing accuracy of the Multiprobe was poor with volumes below 5 μL.

After extraction, tryptic peptides were lyophilized and then resuspended in a minimal volume prior to MALDI plate spotting. A Symbiot I liquid handling robot capable of accurately dispensing 0.5 μL volumes was invaluable for automated spotting of the compact 192-well MALDI plates used specifically for automated MS acquisitions. However, this robot is configured only for spotting MALDI plates from Applied Biosystems. External standards were also spotted adjacent to each sample for optimal calibration and mass accuracy. Alternatively, autolytic tryptic peptides could be used as internal calibrants, however, these peptides frequently are not observed and have a narrow mass range.

After MALDI plate spotting, mass spectra were acquired using the automated acquisition method within Data Explorer software running the Voyager MALDI-TOF. Criteria for accepting or rejecting a mass spectral data file can also be specified here to reduce the accumulation of spectra files with poor or noisy signals. Out of 96 protein spots prepared and analyzed for peptide

Table 1
Overview of robotic instrumentation for automated protein identification from 2-D gels

Model	Manufacturer	Comments (see footnote for list of abbreviations)
Spot excision		
Ettan Picker	Amersham	C, P, 384 spots in single run, gel immersed during spot excision
GelPix	Genetix	C, H, I, P, 8-pin cutting head, HEPA-filtered environment, 1440 spots in single run, 16-bit CCD imaging
ProPic	Genomic Solutions	C, H, I, P, 768 spots in single run, 12-bit CCD camera, also sold as ProXCISION (Perkin–Elmer)
Proteineer spII	Bruker	C, I, P, 384 spots in single run, flatbed scanner imaging
Spot cutter	Bio-Rad	C, I, 96 spots in single run, 12-bit CCD imaging, configurable only with PDQuest (Bio-Rad) analysis software
In-gel digestion liquid handling and/or MALDI plate spotting		
Ettan Digester	Amersham	D, 384 samples in single run, in-gel digestion application only
Ettan Spotter	Amersham	M, 384 samples in single run, MALDI plate spotting application only
Multiprobe II	Packard/ Perkin–Elmer	V, G, D, M, Z, also sold as MassPrep (Micromass, Waters, Bio-Rad), available with automated shaker, heating tiles and 'plate hotel' for processing up to six sample plates
Pro-Prep	Genomic Solutions	C, D, M, Z, HEPA-filtered environment, 384 samples in single run
Proteineer dp	Bruker	D, M, Peltier sample cooling, 384 samples in single run
Symbiot I	Applied Biosystems	M, Z, 288 samples in single run, MALDI plate spotting application only, spots Applied Biosystem MALDI plates only
Hybrid, multi-function robots		
Bullseye	Marsh Bioproducts	C, I, D, M, E, P, 1920 samples in single run, also excises from PVDF
2D iDX	Leap Technologies	C, I, D, M, E, 192 samples in single run, HEPA filtered environment, 8-bit CCD imaging, active monitoring of gel excision
Xcise	Shimadzu Biotech	C, H, I, D, M, Z, E, 1.2 mm fixed excision pin, 8-bit flatbed scanner, 384 samples in single run, does not image fluorescent-stained gels

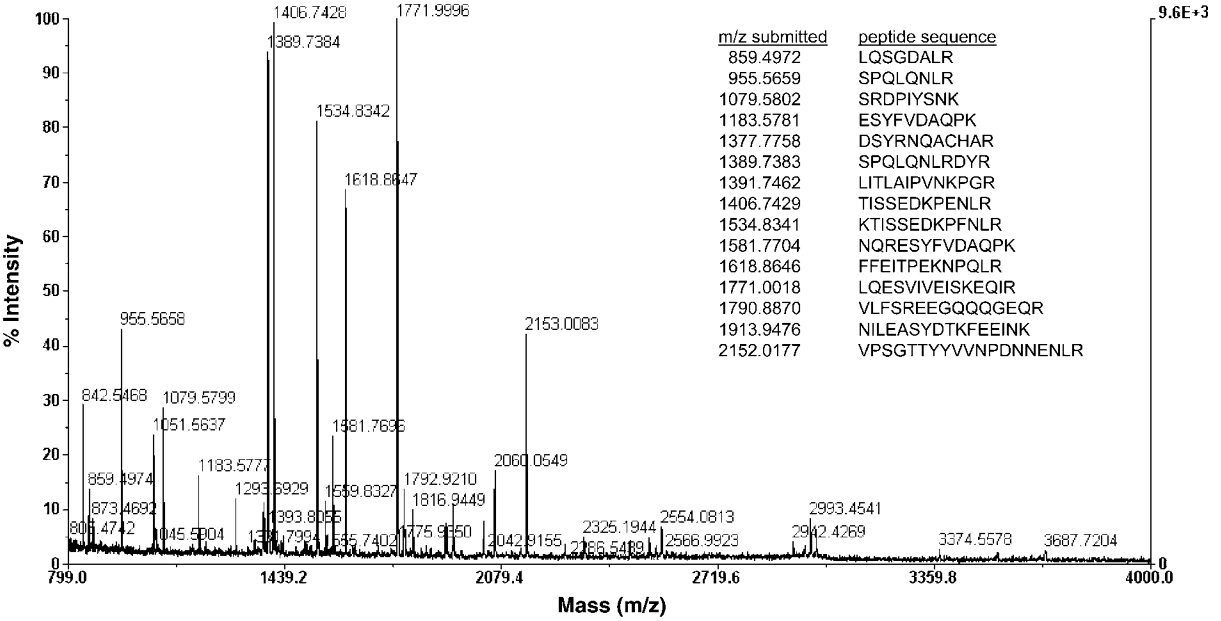
Robotic instrumentation is divided into spot excision, liquid handling and multi-function, hybrid instrumentation. The specific applications each of the hybrid instruments can perform is noted. Number of gel plugs that can be excised and arrayed from the spot-cutting robots in a single, unattended run is also noted.

Abbreviations: C, closed operating environment; H, automated irrigation or humidity control; I, integrated imaging platform and analysis software; P, variable excision pin size; V, configured for filter-bottom plates and vacuum evacuation protocol; G, gripper arm for multiplate processing; E, spot excision function; D, in-gel digestion application; M, MALDI plate spotting function; Z, ZipTip^{TM} microbed C18 desalting application.

mass fingerprinting 62 yielded quality mass spectra, defined as a minimum of fifteen non-isotopic ions of m/z greater than 900 and greater than 10% base peak (for representative spectra see Fig 4). After baseline correction, noise removal, peak deisotoping and threshold adjustment monoisotopic ions from these spectra were submitted to Protein Prospector MS-FIT and protein assignments were made (Fig. 4). Assignments were evaluated based upon molecular weight score (MOWSE), number of peptides matched, peptide mass difference, protein coverage and agreement with molecular weight and pI of intact protein observed by 2-D electrophoresis.

3.3. Protein identification by peptide mass fingerprinting is enhanced by querying contiguous cDNA sequences in UniGene datasets

After processing the 62 mass spectral data files, the peptide masses for each sample were searched against the NCBI non-redundant database (NCBInr) and resulted in an identification frequency lower than 20%. This low rate was attributed to the limited number of complete soybean genes or cDNAs in the NCBInr database. Since the number of legume (soybean and Medicago) ESTs is currently approaching 550,000, the comprehensive EST database (dbEST) was downloaded and queried in addition to the individual legume EST datasets. A preliminary query of dbEST with five samples required 10 h of data processing and attempts to process all 62 samples in batch format resulted in system failure midway through the processing. The preliminary output results from dbEST indicated the high level of false positive candidates would confound data interpretation. In contrast, the soybean EST dataset improved the identification rate to 34% (33/96) which was an improvement over NCBInr and demonstrated the utility of ESTs for soybean proteomic investigations. Finally, querying the soybean UniGene dataset resulted in an overall identification rate of 46% (44/96). Although this percentage appears modest, since only 62 samples met the mass spectra criteria the true identification frequency of this searching method was 71% (44/62). In general, the 34 samples that did not yield quality spectra



m/z submitted	peptide sequence
859.4972	LQSGDALR
955.5659	SPQLQNLR
1079.5802	SRDPIYSNK
1183.5781	ESYFVDAQPK
1377.7758	DSYRNQACHAR
1389.7383	SPQLQNLRDYR
1391.7462	LITLAIPVNKPGR
1406.7429	TISSEDKPENLR
1534.8341	KTISSEDKPFNLR
1581.7704	NQRESYFVDAQPK
1618.8646	FFEITPEKNPQLR
1771.0018	LQESVIVEISKEQIR
1790.8870	VLFSREEGQQQGEQR
1913.9476	NILEASYDTKFEEINK
2152.0177	VPSGTTYYVVNPDNNENLR

Fig. 4. Mass spectrum obtained for tryptic peptides eluted from 2-D gel spot 5. After baseline correction, background subtraction and peak de-isotoping 29 ions were submitted to Protein Prospector. Fifteen of the submitted ions were matched to theoretical tryptic peptides from β-conglycinin α-subunit; the sequence of those peptides are shown next to the mass of the monoisotopic, singly charged ions.

were derived from low abundance or low molecular weight protein spots, both of which reduce the number of diagnostic peptides. Furthermore, the excision pins (1.4 mm diameter) used for spot cutting were generally 1–5 mm smaller in diameter than the CBB-stained spots and therefore retrieved only 5–34% of each protein spot, based upon area. After discussing this with the manufacturer, a 3 mm diameter excision pin was recently produced. Use of the 3 mm diameter excision pin in a recent investigation of the Arabidopsis proteome improved the sensitivity range for automated PMF analyses indicating sample amount is one important factor (data not shown).

Although the utility of public EST databases for PMF investigations was previously noted for maize and *Medicago truncatula* (Porubleva et al., 2001; Watson et al., 2003), use of UniGene datasets and improved protein identification compared with raw EST data was not discussed. The short length and poor sequence quality of raw EST entries would predictably be a hindrance to PMF investigations. It was noted previously that assembly of these large EST datasets into non-redundant contigs could ameliorate these problems and make this vast resource more useful for proteomics investigators (Lisacek et al., 2001). Moreover, abridgment of voluminous and redundant ESTs libraries into unique cDNA contigs drastically reduces the processing time during database mining; an average UniGene search requires 2 min of processing time. For these reasons we have found the library of UniGene databases invaluable for proteome investigations with organisms lacking annotated genome information.

3.4. Seed storage proteins represent over 60% of total soybean seed protein

The two major storage proteins within soybean seed are multimeric and have been characterized biochemically by their sedimentation coefficient. The 7S globulins are comprised of β-conglycinin subunits while the 11S globulins are comprised of glycinin proteins. Of the 44 identified soybean seed proteins identified in this investigation 17 were assigned to these two classes of seed storage proteins (Table 2). Multiple isoelectric or mass species were observed for each of these storage proteins. Since the conglycinin and glycinin classes of storage proteins belong to multigene families, genetic redundancy is one possible explanation for the multiple isoelectric species observed within these protein classes (Schuler et al., 1982a; Nielsen et al., 1989; Harada et al., 1989).

The β-conglycinin class of 7S seed storage proteins is a large multigene family which codes for two mRNA classes of 2.5 and 1.7 kb (Goldberg et al., 1981; Schuler et al., 1982a,b; Harada et al., 1989). The 2.5 kb mRNA encodes two distinct protein subunits α and α′ while the 1.7 kb mRNA encodes the β-subunit (Beachy et al., 1981; Schuler et al., 1982b). These subunits interact to form trimers of differing compositions (Thanh and Shibasaki, 1976b). Automated PMF identification from the 2-D gel in Fig. 2 indicated the abundant 65 kDa protein (spot 8), representing approximately 10% of total protein, corresponded to the α-subunit of β-conglycinin (Table 2). The 70 kDa (spot 4) and 55 kDa (spot 18) proteins were both identified as the α′-subunit of

Table 2
Identities of protein spots from two-dimensional gel analysis of total soybean seed proteins (Fig. 2)

Spot no.	UniGene contig	Protein (μg)	MOWSE score	p*I*	M_r	Identity	Genbank accession	Protein
4	S5146289	179	1.54E + 09	5.4	69.7	76%	AB030838	β-Conglycinin α′-subunit
8	S5146288	99.5	2.81E + 08	4.8	65.1	72%	AB030839	β-Conglycinin α-subunit
11	S5146687	1.46	7.83E + 03	6.5	62.5	81%	Q04672	Sucrose binding protein precursor
13	S5146687	3.00	3.38E + 05	6.4	62.2	81%	Q04672	Sucrose binding protein precursor
14	S5146410	0.38	2.20E + 06	6.3	59.4	85%	AF191299	Sucrose binding protein homolog S-64
18	S5146289	23.8	1.36E + 07	5.3	54.8	76%	AB030838	β-Conglycinin α′-subunit
19	S5146687	0.44	1.09E + 05	6.2	53.5	81%	Q04672	Sucrose-binding protein precursor
20	S5146410	0.79	2.57E + 07	6.6	51.6	85%	AF191299	Sucrose binding protein homolog S-64
21	S5146410	1.46	2.10E + 04	6.3	50.9	85%	AF191299	Sucrose binding protein homolog S-64
25	S5146286	5.43	1.20E + 07	6.6	46.5	95%	AB030841	β-Conglycinin β-subunit
26	S5146286	5.74	1.45E + 09	6.3	45.7	95%	AB030841	β-Conglycinin β-subunit
27	S5146838	0.73	1.18E + 03	7.1	44.9	22%	P04929	Histidine rich glycoprotein precursor
28	S5146286	75.8	3.68E + 08	5.8	44.9	95%	AB030841	β-Conglycinin β-subunit
29	S4875299	0.91	1.86E + 05	7.0	44.7	82%	AY094423	F7F1.19 Hypothetical protein
34	S5146157	8.66	1.05E + 03	7.1	39.5	96%	AF532629	Alcohol dehydrogenase 1
37	S5128976	2.09	4.85E + 02	7.5	38.8	45%	AY080719	Unknown protein
38	S5146653	16.9	1.03E + 04	4.8	38.7	79%	AB000168	Glycinin
39	S5146653	19.7	1.03E + 04	5.0	38.0	79%	AB000168	Glycinin
42	S5146374	4.01	2.42E + 02	7.3	34.7	96%	AF169018	Seed maturation protein
45	S5146746	129	5.91E + 03	5.0	33.6	86%	P04405	Glycinin G2 precursor
47	S5146745	16.0	7.60E + 07	5.2	33.5	88%	P04776	Glycinin G1 precursor
54	S4891789	18.7	1.87E + 04	5.1	31.2	88%	Q42134	20S proteasome alpha subunit E2
55	S5064826	2.99	2.43E + 03	6.9	30.7	46%	AY084581	Hypothetical protein
56	S5011552	3.35	3.16E + 03	5.1	30.0	76%	AJ276270	Hypothetical protein
57	S5146421	2.18	1.95E + 03	4.6	29.8	83%	AF116755	Seed maturation protein PM26
58	S4883917	0.74	2.68E + 03	5.9	29.8	91%	AB086039	Farnesyl pyrophosphate synthase
60	S5146755	8.70	6.48E + 05	9.1	29.3	90%	P13917	Basic 7S globulin precursor
65	S4913497	0.90	3.58E + 03	5.0	24.3	82%	AF307152	Diphosphonucleotide phosphatase
68	S4881345	0.76	1.81E + 02	5.7	23.6	69%	P19594	2S albumin precursor
70	S5096997	0.67	1.08E + 04	5.6	23.2	95%	AF243368	Glutathione *S*-transferase GST 13
73	S5146746	10.7	8.80E + 04	7.7	21.6	86%	P04405	Glycinin G2 precursor
74	S4862087	0.10	4.60E + 02	7.0	21.5	90%	AL096860	40S ribosomal protein S20-like protein
75	S5146744	3.45	1.40E + 05	9.3	21.3	82%	P02858	Glycinin G4 precursor
76	S5146746	60.0	9.56E + 05	8.6	21.2	86%	P04405	Glycinin G2 precursor
77	S5146745	20.6	9.74E + 01	7.7	21.1	88%	P04776	Glycinin G1 precursor
79	S4876999	0.72	1.30E + 02	7.0	20.9	94%	Q9SXU1	Proteasome subunit alpha type 7
81	S5146653	5.60	4.87E + 03	9.5	20.8	79%	AB000168	Glycinin
82	S5146745	4.85	1.05E + 04	7.7	20.7	88%	P04776	Glycinin G1 precursor
85	S4865740	0.23	6.40E + 02	6.1	20.2	82%	AB005232	MBG8.21 translation initiation factor like
88	S4891239	1.92	4.95E + 02	6.6	19.0	44%	AF411801	Hypothetical protein
89	S5146661	1.68	2.96E + 03	5.6	18.6	97%	P08170	Seed lipoxygenase-1
90	S4910645	5.25	2.95E + 03	6.4	18.6	62%	AY085867	Hyloglucan endo-1, 4-β-D-glucanase
91	S4882919	1.36	5.42E + 02	6.9	18.6	41%	AC005489	F14N23.2 hypothetical protein
94	S5146744	25.0	3.95E + 02	5.9	14.4	82%	P02858	Glycinin G4 precursor

Glycine max UniGene cDNA contig number for each highest scoring assignment from Protein Prospector is noted. Protein amount for each spot was determined using a series of protein standards resolved in the second dimension of the gel and quantified using Phoretix 2D Advanced. Each spot was normalized to spot 76 which was determined to be 60 μg. Molecular weight search (MOWSE) scores for each entry are noted. In general, a score greater than 100 is significant. UniGene contig DNA sequence number from the *Glycine max* database is noted. Each UniGene entry was BLAST searched against the NCBI non-redundant database to determine the highest scoring match. Genbank accession numbers, protein identifier and percent identity to the UniGene entry are noted.

β-conglycinin and together represented 21% of the total protein. The abundant 45–47 kDa isoelectric series of spots (25, 26, 28) were separately identified as β-subunits of β-conglycinin. These assignments are in agreement with previous SDS–PAGE analysis of β-conglycinin subunits (Tumer et al., 1981) and a soybean seed proteomics investigation published during the writing of this manuscript (Herman et al., 2003). In addition to the β-conglycinin subunits, protein assignments to abundant proteins such as glycinin subunits (spots 38, 39, 45, 47, 73, 75–77, 81, 82, 94), basic 7S albumin (spot 68), sucrose binding proteins (spots 11, 13, 14, 19–21), seed maturation protein (AF169018, spot 42) and alcohol dehydrogenase (spot 34) were each in agreement with the recent report by Herman et al. (2003). Although the spot location and identity of the aforementioned proteins were confirmed with the report by Herman et al. most of the remaining, lower abundance protein spots could not be unequivocally aligned with the proteome map in that report. This could be due the fact that different soybean varieties as well as protein isolation and electrophoresis protocols were employed. Comparisons could also be confounded by the fact that the proteome reference map in the previous investigation was a computational composite image rather than a true 2-D gel image. Nevertheless, the previous investigation has serendipitously validated the automated PMF workflow discussed here. This is particularly notable since the protein identification methods were different.

Although protein separation by 2-D electrophoresis is susceptible to reproducibility problems this technique is valuable for discovering proteins that are modified by the numerous post-translational modifications that alter protein isoelectric point. For instance, the β-subunits to β-conglycinin and sucrose binding proteins were each comprised of at least five isoelectric spots with similar, respective masses (Figs. 1 and 2). This unusually high number of isoelectric species is suggestive of post-translational modification. Indeed, the β-conglycinin subunits are well known glycoproteins (Thanh and Shibasaki, 1976b; Thanh and Shibasaki, 1977) and *N*-glycosylation is a modification which is capable of causing a change in isoelectric point (Loster and Kannicht, 2002).

In addition to the 7S globulin, another major storage protein within soybean seeds is the 11S globulin composed of glycinin polypeptides (Badley et al., 1975; Barton et al., 1982). Two groups of glycinin storage proteins were identified on the proteome map for soybean seed. One group of proteins was approximately 35 kDa with acidic pI values (spots 43, 45, 47) while the second set was 21 kDa and strongly basic (spots 73, 75, 76, 77, 81, 82). These protein spots correspond with the previous assignment of six acidic subunits of 36–40 kDa and six basic subunits of 20–22 kDa (Badley et al., 1975; Barton et al., 1982). Two higher molecular weight, acidic glycinins (spots 38, 39; 39 kDa) were also observed and possibly represent the previously identified A4 subunits of glycinin. (Nielsen et al., 1989; Nielsen, 1996). These two proteins comprised approximately 15% of the acidic glycinins and were distinct from the predominant 35 kDa forms.

Additional seed storage proteins that were identified include the 29 kDa subunit (spot 60) to a 7S globulin glycoprotein termed Bg (Watanabe and Hirano, 1994) and the 2S albumin precursor (spot 68) previously characterized as a methionine-rich cotyledon protein (Revilleza et al., 1996). Similar to 11S glycinins the mature Bg globulin is derived from proteolytic processing of a single precursor polypeptide. Specifically, Bg glycoprotein is post-translationally processed into a basic 29 kDa α-subunit and a 16 kDa β-subunit which are bound by disulfide bridging (Kagawa et al., 1987; Watanabe and Hirano, 1994). The 29 kDa α-subunit was positively identified (spot 60, Fig. 2) and represented approximately 1.2% of the total seed protein; a minor contribution compared to the β-conglycinin and glycinin storage proteins.

3.5. An abundant isoelectric series of 62 kDa proteins are sucrose binding proteins

Another abundant protein in soybean seed is sucrose binding protein (SBP; spots 11, 13, 14, 19–21) which represents approximately 1% of total seed protein. These proteins are believed to be important for source-sink relations as they are responsible for binding sucrose as it is deposited in the cells of developing cotyledons as well as mesophyll cells of young sink leaves and the companion cells of mature phloem (Grimes et al., 1992). These proteins also bear sequence and structural similarity to the vicilin-like seed storage proteins (Overvoorde et al., 1997). However, SBP has been characterized as a peripheral membrane protein localized to the external leaflet of the plasma membrane (Overvoorde and Grimes, 1994). Although SBPs have been previously characterized, the abundance of these proteins in mature soybean seed and observation of multiple isoelectric species has, to our knowledge, not been reported. It was noted recently that two genes code for SBPs in the soybean genome (Contim et al., 2003), although at least six different protein spots were identified as SBPs (Table 2). These data suggest post-translational processing or modification might be occurring with these proteins.

Alcohol dehydrogenase (ADH) also accumulated to nearly 1% of the total protein in mature soybean seed (spot 34). However, unlike SBPs only one protein species was observed for ADH. The abundance of ADH in soybean seed is not surprising as the low oxygen environment within embryos of legumes induces fermentative metabolism and in particular ADH (Rolletschek et al., 2003). Although embryogenic photosynthesis provides oxygen for aerobic metabolism later in development the amount of ADH activity is nevertheless substantial in maturing embryos (Rolletschek et al., 2003). Alternatively, a high basal level of ADH may be necessary for successful germination under low oxygen conditions as observed in other, non-legumes (Conley et al., 1999; Fukao et al., 2003). Identification of two seed maturation proteins (spots 42 and 57) and a seed lipoxygenase (spot 89) is also consistent with the source material for this proteomic investigation. However, the low molecular weight of spot 89 is inconsistent with the predicted masses of lipoxygenases (greater than 80 kDa) and therefore may represent a proteolytic product.

Even though the majority of proteins identified in this preliminary survey of soybean seed could be explained in the context of previous research in seed biology, the biological function of at least eight assigned protein spots was not evident. For instance, four protein spots (29, 55, 56, 88) were each assigned as hypothetical proteins, of which the function is not yet known. Also, identification of diphosphonucleotide phosphatase, glutathione *S*-transferase, and two subunits to the proteasome protein degradation complex was unexpected. Although the catalytic function of these enzymes is well known, the role of these proteins in seed biology is not known and will thus require further confirmation and experimentation.

4. Concluding remarks

The high amount of sample handling and processing required for PMF protein identification from 2-D gels has placed a premium on laboratory automation for high-throughput proteomics. An automated workflow also reduces the level of sample contamination and substantially increases the throughput of sample processing. The sample throughput of automated PMF is at least three-fold greater than electrospray ionization-tandem MS and is not hindered by sample contamination arising from incomplete column elutions from 'front-end' capillary liquid chromatography.

A preliminary survey of the soybean seed proteome revealed the major obstacle towards developing an in-depth 2-D gel proteome map is the broad dynamic range of protein expression, ascribed to the preponderance of seed storage proteins. Development of a simple technique to remove the 7S and 11S globulins would be a significant achievement and facilitate greater proteome coverage. Despite this inherent problem, over one hundred proteins was resolved and quantitated by 2-D electrophoresis. An automated PMF approach towards protein identification resulted in the assignment of 44 proteins, including many previously unknown proteins. Identification of the remaining proteins will likely require the retrieval of more protein for PMF or liquid chromatography tandem MS acquisitions on the low abundance proteins. Continued research towards the development of a soybean seed proteome map will be useful for rapid comparison of soybean cultivars, mutants and transgenics. Future investigations into seed compositional analysis, seed-filling and seed physiology will also benefit from a detailed and quantitative proteome reference map of soybean seed.

Acknowledgements

The authors acknowledge Dr. Hari Krishnan for providing seed material and critical reading of this manuscript. This research was supported by NSF grant DBI-0332418 (JJT) and USDA-CSREES grant 2001-35318-12971 (JJT).

References

Badley, R.A., Atkinson, D., Hauser, H., Oldani, D., Green, J.P., Stubb, J.M., 1975. The structure, physical and chemical properties of the soybean protein glycinin. Biochim. Biophys. Acta 412, 214–228.

Barton, K.A., Thompson, J.F., Madison, J.T., Rosenthal, R., Jarvis, N.P., Beachy, R.N., 1982. The biosynthesis and processing of high molecular weight precursors of soybean glycinin subunits. J. Biol. Chem. 257, 6089–6095.

Beachy, R.N., Jarvis, N.P., Barton, K.A., 1981. Biosynthesis of subunits of the soybean 7S storage protein. J. Mol. Appl. Genet. 1, 19–27.

Conley, T.R., Peng, H.P., Shih, M.C., 1999. Mutations affecting induction of glycolytic and fermentative genes during germination and environmental stresses in Arabidopsis. Plant Physiol. 119, 599–608.

Contim, L.A., Waclawovsky, A.J., Delu-Filho, N., Pirovani, C.P., Clarindo, W.R., Loureiro, M.E., Carvalho, C.R., Fontes, E.P., 2003. The soybean sucrose binding protein gene family: genomic organization, gene copy number and tissue-specific expression of the SBP2 promoter. J. Exp. Bot. 54, 2643–2653.

Cottrell, J.S., 1994. Protein identification by peptide mass fingerprinting. Pept. Res. 7, 115–124.

Fukao, T., Kennedy, R.A., Yamasue, Y., Rumpho, M.E., 2003. Genetic and biochemical analysis of anaerobically-induced enzymes during seed germination of *Echinochloa crus-galli* varieties tolerant and intolerant of anoxia. J. Exp. Bot. 54, 1421–1429.

Goldberg, R.B., Hoschek, G., Ditta, G.S., Breidenbach, R.W., 1981. Developmental regulation of cloned superabundant embryo mRNAs in soybean. Dev. Biol. 83, 218–231.

Grimes, H.D., Overvoorde, P.J., Ripp, K., Franceschi, V.R., Hitz, W.D., 1992. A 62-kD sucrose binding protein is expressed and localized in tissues actively engaged in sucrose transport. Plant Cell 4, 1561–1574.

Harada, J.J., Barker, S.J., Goldberg, R.B., 1989. Soybean beta-conglycinin genes are clustered in several DNA regions and are regulated by transcriptional and posttranscriptional processes. Plant Cell 1, 415–425.

Herman, E.M., Helm, R.M., Jung, R., Kinney, A.J., 2003. Genetic modification removes an immunodominant allergen from soybean. Plant Physiol. 132, 36–43.

Herman, E.M., Larkins, B.A., 1999. Protein storage bodies and vacuoles. Plant Cell 11, 601–613.

Hu, B., Esen, A., 1981. Heterogeneity of soybean seed proteins: one-dimensional electrophoretic profiles of six different solubility fractions. J. Agric. Food Chem. 29, 497–501.

Hu, B., Esen, A., 1982. Heterogeneity of soybean proteins: two-dimensional electrophoretic maps of three solubility fractions. J. Agric. Food Chem. 30, 21–25.

Hurkman, W.J., Tanaka, C.K., 1986. Solubilization of plant membrane proteins for analysis by two-dimensional gel electrophoresis. Plant Physiol. 81, 802–806.

Kagawa, H., Yamauchi, F., Hirano, H., 1987. Soybean basic 7S globulin represents a protein widely distributed in legume species. FEBS Lett. 226, 145–149.

Lisacek, F.C., Traini, M.D., Sexton, D., Harry, J.L., Wilkins, M.R., 2001. Strategy for protein isoform identification from expressed sequence tags and its application to peptide mass fingerprinting. Proteomics 1, 186–193.

Loster, K., Kannicht, C., 2002. 2D-Electrophoresis: detection of glycosylation and influence on spot pattern. In: Christoph

Kannicht (Ed.), Postranslational Modification of Proteins, Humana Press, pp. 301–316.

Mann, M., Hendrickson, R.C., Pandey, A., 2001. Analysis of proteins and proteomes by mass spectrometry. Annu. Rev. Biochem. 70, 437–473.

Nielsen, N.C., Dickinson, C.D., Cho, T.-J., Thanh, V.H., Scallon, B.J., Fischer, R.L., Sims, T.L., Drews, G.N., Goldberg, R.B., 1989. Characterization of the glycinin gene family in soybean. Plant Cell 1, 313–328.

Nielsen, N.C., 1996. Soybean seed composition. In: Verma, D.P.S., Shoemaker, R.C. (Eds.), Soybean: Genetics, Molecular Biology and Biotechnology. CAB, Wallingford, pp. 127–163.

Overvoorde, P.J., Grimes, H.D., 1994. Topographical analysis of the plasma membrane-associated sucrose binding protein from soybean. J. Biol. Chem. 269, 15154–15161.

Overvoorde, P.J., Chao, W.S., Grimes, H.D., 1997. A plasma membrane sucrose-binding protein that mediates sucrose uptake shares structural and sequence similarity with seed storage proteins but remains functionally distinct. J. Biol. Chem. 272, 15898–15904.

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

Pieper, R., Su, Q., Gatlin, C.L., Huang, S., Anderson, N.L., Steiner, S., 2003. Multi-component immunoaffinity subtraction chromatogtography: an innovative step towards a comprehensive survey of the human plasma proteome. Proteomics 3, 422–432.

Pontius, J.U., Wagner, L., Schuler, G.D., 2003. UniGene: a unified view of the transcriptome. In: The NCBI Handbook, Bethesda (MD): National Center for Biotechnology Information.

Porubleva, L., Vander Velden, K., Kothari, S., Oliver, D.J., Chitnis, P.R., 2001. The proteome of maize leaves: use of gene sequences and expressed sequence tag data for identification of proteins with peptide mass fingerprints. Electrophoresis 22, 1724–1738.

Revilleza, M.J., Galvez, A.F., Krenz, D.C., de Lumen, B.O., 1996. An 8 kDa methionine-rich protein (MRP) from soybean (*Glycine max*) cotyledon: identification, purification and N-terminal sequence. J. Agric. Food Chem. 44, 2930–2935.

Rolletschek, H., Weber, H., Borisjuk, L., 2003. Energy status and its control on embryogenesis of legumes. Embryo photosynthesis contributes to oxygen supply and is coupled to biosynthetic fluxes. Plant Physiol. 132, 1196–1206.

Schuler, M.A., Schmitt, E.S., Beachy, R.N., 1982a. Closely related families of genes code for the alpha and alpha' subunits of the soybean 7S storage protein complex. Nucl. Acid Res. 10, 8225–8244.

Schuler, M.A., Ladin, B.F., Fryer, G., Pollaco, J.G., Beachy, R.N., 1982b. Structural sequences are conserved in the genes coding for the α', α and β subunits of the soybean seed storage protein and other seed proteins. Nucl. Acids. Res. 10, 8245–8260.

Steinberg, T.H., Chernokalskaya, E., Berggren, K., Lopez, M.F., Diwu, Z., Haugland, R.P, Patton, W.F., 2000. Ultrasensitive fluorescent protein detection in isoelectric focusing gels using a ruthenium metal chelate stain. Electrophoresis 21, 486–496.

Thanh, V.H., Shibasaki, K., 1976a. Major proteins of soybean seeds. A straightforward fractionation and their characterization. J. Agric. Food Chem. 24, 1117–1121.

Thanh, V.H., Shibasaki, K., 1976b. Heterogeneity of beta-conglycinin. Biochim. Biophys. Acta 439, 326–338.

Thanh, V.H., Shibasaki, K., 1977. Beta-conglycinin from soybean proteins. Isolation and immunological and physicochemical properties of the monomeric forms. Biochim. Biophys. Acta 490, 370–384.

Tumer, N.E., Thanh, V.H., Nielsen, N.C., 1981. Purification and characterization of mRNA from soybean seeds: identification of glycinin and β-conglycinin precursors. J. Biol. Chem. 256, 8756–8760.

Watanabe, Y., Hirano, H., 1994. Nucleotide sequence of the basic 7S globulin gene from soybean. Plant Physiol. 105, 1019–1020.

Watson, B.S., Asirvatham, V.S., Wang, L., Sumner, L.W., 2003. Mapping the proteome of barrel medic (*Medicago truncatula*). Plant Physiol. 131, 1104–1123.

Wilkins, M.R., Sanchez, J.C., Gooley, A.A., Appel, R.D., Humphery-Smith, I., Hochstrasser, D.F., Williams, K.L., 1995. Progress with proteome projects: why all proteins expressed by a genome should be identified and how to do it. Biotech. Gene Eng. Rev. 13, 19–50.

Yates III, J.R., Speicher, S., Griffin, P.R., Hunkapiller, T., 1993. Peptide mass maps: a highly informative approach to protein identification. Anal. Biochem. 214, 397–408.