Plant Omics Journal POJ

POJ 4(5):257-263 (2011) ISSN:1836-3644

Improved protein identification using a species-specific protein/peptide database derived from expressed sequence tags

Jinhui Chen[1,2], Jisen Shi[2], Dagang Tian[1], Liming Yang[1], Yuming Luo[1,*], Denghua Yin[1], Xiangyang Hu[3]

[1]School of Life Sciences, Huaiyin Normal University, Jiangsu Key Laboratory for Eco-Agricultural Biotechnology around Hongze Lake, Huaian 223300, Jiangsu, China
[2]Key Laboratory of Forest Genetics & Biotechnology, Ministry of Education, Nanjing Forestry University, Nanjing 210037, China
[3]Key Laboratory of Biodiversity and Biogeography, Kunming Institute of Botany, Institute of Tibet Plateau Research at Kunming, Chinese Academy of Science, Kunming 650204, China

*** Corresponding authors: Yuming Luo (yanglm2@yahoo.com.cn)**

Abstract

The use of peptide mass fingerprinting data obtained by matrix-assisted laser desorption/ionisation time-of-flight mass spectrometry in conjunction with protein database searching is a fast, high-throughput method that is widely used to identify proteins in a proteome. The success of a search is limited, however, by the number of proteins represented in the database. When the fully sequenced genome of an organism is not available, cross-species databases are usually used, which can compromise the reliability of the results. Databases containing expressed sequence tag sequences are available and are potentially invaluable resources for proteomic studies. For the study reported herein, we developed a proteomic approach that incorporates a species-specific protein/peptide database constructed from expressed sequence tag sequences, and we validated this approach by accurately retrieving proteins from this database that matched, on the basis of the masses of their *in silico*–generated tryptic peptides, experimentally isolated proteins derived from a wheat stem proteome. We also compared the results obtained using the wheat database with those obtained using the same peptide mass fingerprints and a cross-species protein database search. Given the reliability of the results and the improved scoring obtained with the wheat database study in comparison with the cross-species database study, species-specific databases derived from expressed sequence tags may replace cross-species databases for proteomic studies involving organisms for which a completely sequenced genome is unavailable.

Keywords: Cross-species protein database; Peptide mass fingerprinting; Protein identification; Proteomics.
Abbreviations: 2-DE, two-dimensional gel electrophoresis; EST, Expressed sequence tags; MALDI-TOF MS, matrix-assisted laser desorption/ionisation-time of flight mass spectrometry; NCBInr protein database, National Center for Biotechnology Information nonredundant protein database; PMF, peptide mass fingerprinting.

Introduction

Large-scale protein expression studies, i.e., proteomic studies, are a powerful way to characterise the function and regulation of genes (Yates, 2000; Aebersold and Mann, 2003). Using two-dimensional gel electrophoresis (2-DE), mass spectral characterisation of peptide mass fingerprints (PMFs) derived from tryptic digests of electrophoretically isolated proteins, and database searching, proteomic studies can address various biological questions and have become the standard means for large-scale protein identification (Mallick and Kuster, 2010). With the increasing number of genomic databases and the number of sequences within the databases that have not been assigned to known proteins, large-scale protein identification has become more important and more widely attempted. The identification of experimentally isolated proteins by mass spectrometry (MS) analysis and/or sequence characterisation of their enzymatically digested peptides relies on bioinformatics software (such as MASCOT, ProFOUND, MS-Fit) to retrieve the corresponding proteins from an appropriate database. After matching the experimental mass spectral data with those in a database, the corresponding matched proteins are sorted using a scoring algorithm, and the one with the greatest score is often considered to be the experimentally isolated protein (Palagi et al. 2009). Among the available proteomic methods, PMF by MS combined with database searching is a fast, high-throughout method for protein identification and is widely used. However, the success of this method largely depends on the presence of the corresponding proteins in the database that is searched. For an organism with a genome that has been fully sequenced and annotated, e.g., *Arabidopsis thaliana* and rice, this method has been successfully used to retrieve targeted proteins from the corresponding genome and/or protein databases (Rajjou et al., 2006; Higashi et al., 2006; Yang et al., 2007). For most organisms, however, the genome has not been completely sequenced, and only a relatively small number of protein sequences are available in protein databases. Therefore, experimental PMF data have usually been used to search a cross-species database, e.g., the NCBInr protein database (Yahata et al., 2005; Dong et al.,

2006; Kamal et al., 2009; 2010; Shin et al., 2010; Scippa et al., 2010), and the accuracy of the results can vary greatly depending on the organism studied. When a cross-species protein database and PMF are used, proteins are identified according to their masses, which reflect the similarity or conservation with sequences in the database (Wright et al., 2010). However, cross-species protein identification is not the same as searching for the full or partial sequences of the targeted proteins, which compensates for potential single-nucleotide polymorphisms among homologues. Because proteomic methods rely on the masses of relatively small peptides, even a small number of nonidentical residues in peptides from homologues may be reflected as large differences in peptide masses. Therefore, matching peptides in a database to their experimental counterparts is difficult, and misidentification results if the PMF data are affected by polymorphism (Wright et al., 2010). Consequently when cross-species databases are used to identify proteins in conjunction with the masses of experimentally obtained peptides, low scores are usually returned and sequence coverage is limited (Mathesius et al. 2002), both of which undermine confidence in the results. Species-specific EST sequence databases have been used when the genome of an organism has not been fully sequenced (Mathesius et al., 2001; Lisacek et al., 2001; Kim et al., 2003; Kwon et al., 2003), and their use can overcome the shortcomings inherent in the use of cross-species genomic and/or protein databases. The results obtained when the same PMF data were used to search species-specific EST and cross-species protein databases have been compared (Porubleva et al., 2001; Mathesius et al., 2002; Kim et al., 2003; Mooney et al., 2004; Grimplet et al., 2005; Huang et al., 2006; Edwards 2007), and the reliability of the protein identifications were greater when the EST databases were used, despite sequence inaccuracies in the ESTs themselves. For example, Watson et al. (2003) reported that when the PMF method was used in conjunction with the NCBInr and SwissProt databases, only 25% of the experimentally obtained proteins were correctly matched to the database proteins even though the mass spectra were of good quality and that the number of correctly identified proteins substantially increased (up to 46%) when the EST database was used. Usually, however, identification of proteins from organisms that have not had their genomes completely sequenced has relied on the use of cross-species databases (Kamal et al., 2009; Peng et al., 2009; Kamal et al., 2010; Shin et al., 2010;) or sometimes in combination with species-specific EST sequence databases (Østergaard et al., 2004; Dong et al., 2006; Laino et al., 2010; Irar et al., 2010; Scippa et al., 2010). Species-specific EST databases have usually not been used in isolation, perhaps because of the uneven quality of the EST data, which impedes the necessary six-fold translation of EST sequences. Without the availability of the full complement of sequences, the extent of sequence coverage, the output score, and accuracy of protein identification are negatively affected. Moreover, it is difficult to distinguish the best candidate protein from other candidates when their scores and sequence coverage are similar (Lisacek et al., 2001; Habermann et al., 2004; Wright et al., 2010). Therefore, incorporating EST sequence information into a proteome investigation for an organism that does not have a completely sequence genome has been challenging. We recently reported a method that uses species-specific EST sequences to improve the efficiency and accuracy of protein identification (Yang et al., 2010). However, this method is time-consuming and requires a large amount of manual labour, and the uneven quality of the EST data makes the identification process difficult. For the work reported herein, we provide a strategy for the construction of a species-specific, EST-derived protein/peptide database that resolves the aforementioned limitations of our previously reported method. The method involves searching the species-specific, EST-derived database with mass spectral PMF data. We also compared the results obtained using this method when a species-specific (wheat) EST-derived database and a cross-species (NCBInr) database were use to determine if, with the use of the species-specific database, the results will be reliable enough for high-throughout proteomic research involving an organism that does not have a fully sequenced genome.

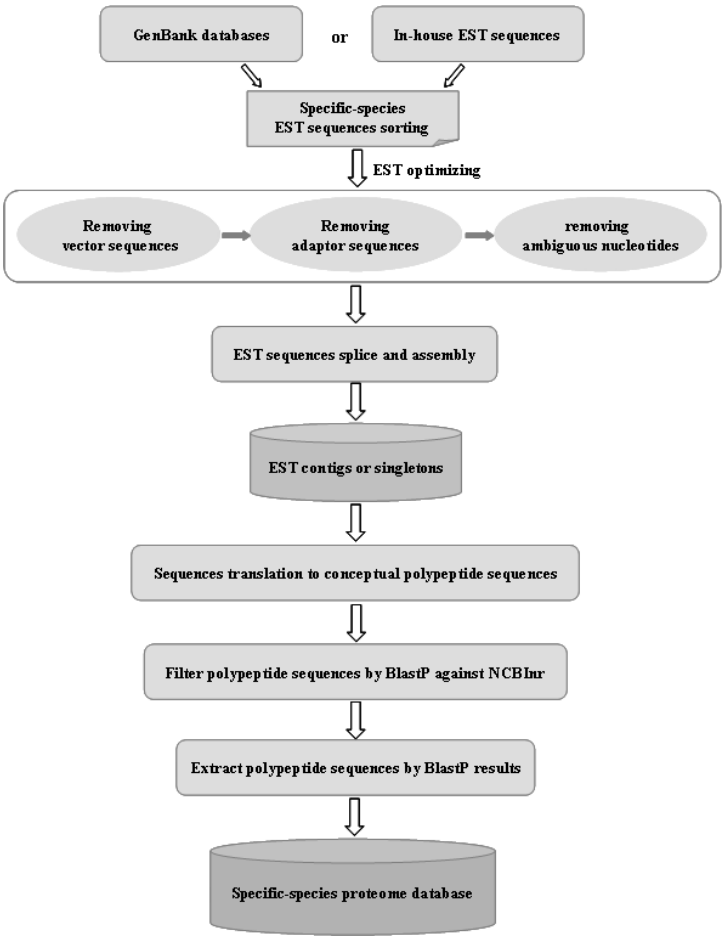

Fig 1. Flowchart showing how a species-specific protein/peptide database is built that can be used in a proteomic study for the identification of proteins from an organism for which a fully sequenced genome is unavailable.

Results and discussion

Construction of a species-specific protein/peptide database

Fig. 1 presents a flow chart diagramming how to construct a

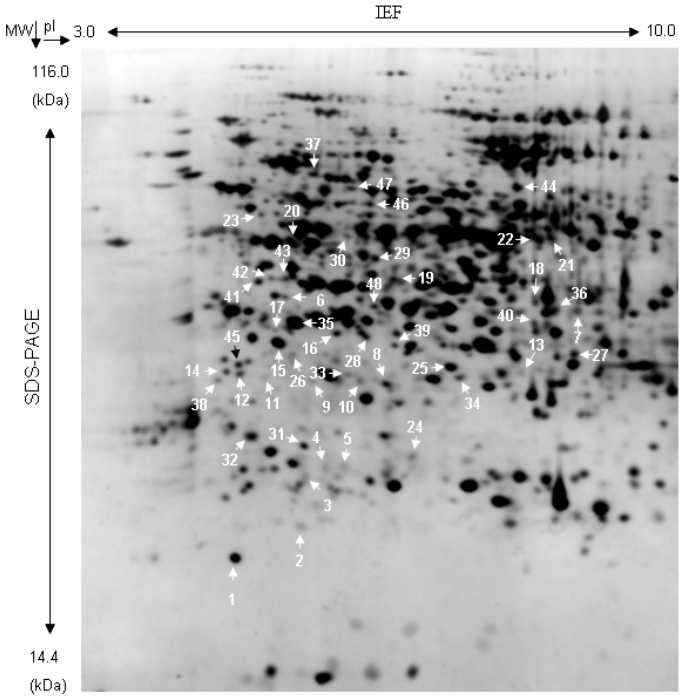

Fig 2. Two-dimensional separation of wheat-stem proteins. First dimension, isoelectric focusing (IEF). Second dimension, separation via SDS-PAGE (12.5% polyacrylamide gel). Arrows identify the 48 protein spots used in this study.

species-specific, EST-derived protein/peptide database, and the steps involved in its construction are as follows.

1. Species-specific ESTs are obtained from the National Center for Biotechnology Information (Bethesda, MD, USA) dbEST database (http://www.ncbi.nlm.nih.gov/dbEST), another public EST database, and/or from an in-house-constructed EST database.
2. Vector sequences are identified using VecScreen (http://www.ncbi.nlm.nih.gov/VecScreen/VecScreen.html; Altschul et al., 1997) and cross_match (http://www.phrap.org) against the UniVec database and then removed from the EST sequence list.
3. Adaptor sequences are removed from the EST sequence list using the cutadapt program (downloaded from http://download.famouswhy.com/cutadapt/).
4. Sequences are processed using trimseq (http://emboss.sourceforge.net/apps/release/5.0/emboss/apps/trimseq.html) to remove ambiguously identified nucleotides found at sequence ends.
5. ESTs are then optimally spliced and assembled into EST contigs or left as singletons according to their sequence similarity (>98%) and 100–bp-overlap length or according to custom-defined parameters using CAP3 (Huang and Madan 1999).
6. The EST contigs and singletons are translated in the 5'→3' and 3'→5' directions and in the three possible reading frames using Transeq program (http://www.ebi.ac.uk/tools/emboss/transeq/; Rice et al., 2000). Then, the six *in silico* translation products serve as queries used to search the NCBInr protein database by BLASTP (http://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&BLAST_PROGRAMS=blastp&PAGE_TYPE=BlastSearch&SHOW_DEFAULTS=on&LINK_LOC=blasthome) to identify candidate gene products for subsequent species-specific protein/peptide database construction. Translated sequences that have no matching results are deleted.
7. All possible polypeptide sequences obtained in the previous step are collected, annotated by querying NCBInr plant protein database, and compiled in FASTA format, and this collection serves as the species-specific protein/peptide database.

Application of a species-specific database for protein identification

To evaluate the efficiency of protein identification when a species-specific database is used, we used proteins from an experimentally obtained wheat-stem proteome and, for comparison, the NCBInr green plant (Viridiplantae) protein database (ftp://ftp.ncbi.nih.gov/) and a species-specific wheat protein/peptide database. A small-scale wheat database was constructed from wheat EST sequences (ftp://ftp.ncbi.nih.gov/) using the strategy described above. The wheat-stem proteome was separated by 2-DE. Fifty protein spots (Fig. 2) were isolated and digested with trypsin. High-quality MALDI-TOF MS PMFs were obtained for all but two of the corresponding proteins. A representative mass spectrum (spot 3) is provided in Fig. 3. The spectra were internally calibrated with two autolytic tryptic peptides (the values for m/z are 842.509 and 2211.104). Each mass list was automatically generated and checked manually against the corresponding spectrum. Mascot program (http://www.matrixscience.com) was used in conjunction with the PMF data from the 48 protein spots to search the NCBInr plant protein database, which is frequently used for plant protein identification, and the wheat database. The search parameters were: a mass tolerance of 50 ppm, a requirement for carbamidomethyl cysteines, allowance for methionine oxidation, allowance for one incomplete trypsin cleavage event per preptide, and the monoisotopic peptide mass (MH^+). The identity of a protein retrieved from a database was considered to be correct if it was the top score in the MASCOT list, had a calculated mass equal to ±20% of the experimentally determined mass, had a calculated pI within ±1 pH units of the experimentally determined pI, and more than 20% of the masses of the *in silico*–generated tryptic peptides from the retrieved protein were matched to those of the experimental peptides. The proteins retrieved for the 48 protein spots from both databases are listed in Table S1. The identities of 25 spots (spots 24–48) were the same when either database was used, whereas the identities of 23 spots (spots 1–23) were dependent on the database used. On average, when the same protein was identified, its sequence coverage and the number of matched peptides were greater when it was retrieved from the wheat database than from the cross-species database. When the cross-species database was used for spots 1–23, their masses and pI values also deviated more from the calculated values than when the wheat database was used (Table S1; Fig. 4B, C). Therefore, for spots 1–23, the corresponding proteins retrieved from the cross-species database were less likely to have been correctly identified than those retrieved from the wheat database, which were verified by ESI-MS/MS analysis. Additionally, the sequence coverage and the number of matched peptides for spots 1–23 were also

greater when the wheat database was used (Table S1; Fig. 4A). For spots 24–48, the searches of the two databases returned identical proteins as noted above. According to the annotations in the NCBInr database, seven protein spots (spots 24–30) matched those in the taxon *Triticum aestivum*, eight (spots 31–37) matched those in the taxon *Hordeum vulgare*, seven (spots 38–44) matched those in the taxon *Oryza sativa*, three (spots 45–47) matched those in the taxon *Zea mays*, and one (spot 48) matched one in the taxon *Populus trichocarpa*. Despite identifying the same proteins for spots 24–48 when either database was used, the sequence coverage and the number of matched peptides were smaller when the cross-species database was used (Table S1; Fig. 4A). Therefore, protein retrieval was probably affected by the polymorphism that existed between the amino acid sequences of the homologous proteins from the different species in the cross-species database, which led to large differences in the masses of the queried and retrieved proteins, lower scores, lower sequence coverage, and smaller numbers of matched peptides; consequently, the mass matching of the *in silico*–generated peptides and the experimental peptide ions was more difficult. We also generated a *Nitraria sibirica* protein/peptide database from an in-house EST database using 454 Sequencing Systems (Roche NimbleGen, USA), and a *Crotalus atrox* protein/peptide database using EST sequences downloaded from http://www.ncbi.nlm.nih.gov/sra/SRP002110 (Gracheva et al., 2010). These species-specific protein/peptide databases are currently being used to identify proteins from their proteomes. Preliminary results suggested that the use of these species-specific databases would produce more reliable results than would searching the NCBInr protein database. To determine if protein identification by species-specific, EST-derived protein/peptide database searching with PMF data is accurate enough for high-throughout proteomics studies that use organisms for which complete genome sequence information is unavailable, eight protein spots from the wheat proteome were digested with trypsin, and the peptides were subjected to electrospray ionisation tandem MS sequencing, which confirmed that the identities of the proteins retrieved from the wheat database using the PMF data were the correct ones (data not shown). For the work reported herein, we developed a method that appears to reliably identify electrophoretically separated proteins from organisms with incompletely sequenced genomes by mass spectral PMF and searching a species-specific protein/peptide database derived from EST sequences. When this strategy was applied to a wheat-stem proteome, the number of matched peptides and the sequence coverage usually increased for a given protein spot in comparison with a search made using the NCBInr database, a cross-species database. With the use of modern DNA sequencing technologies, a large number of nucleotide sequences have been and continue to be determined. Different types of nucleotide sequence databases are becoming available, and EST databases are the fastest growing type of nucleotide database (Neale and Kremer, 2011). ESTs are used to refine the sequences of predicted genes, and such gene sequences can then be used to predict protein sequences and functions. EST sequences represent an enriched set of transcripts as they are derived from expressed genes, and their sequences are clearly worth exploiting in proteomic studies (Mathesius et al., 2002; Kwon et al., 2003; Merlino et al., 2009; Irar et al., 2010; Scippa et al., 2010). The identification of proteins from ESTs and EST contigs will be an interesting and appropriate way to expand the content of protein databases.

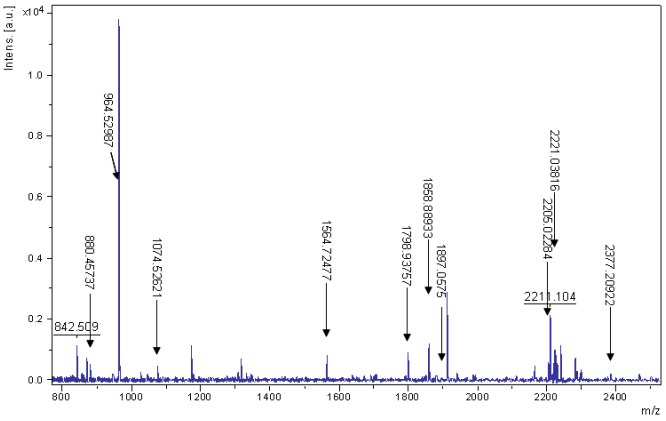

Fig 3. A representative mass spectrum (for spot 3). The peaks for which matching peptides were found in the wheat database are labelled with their *m*/*z* values.

Materials and methods

Wheat growth and protein extraction from wheat stems

The wheat cultivar Yangmai No. 3 was grown in 40 cm (diameter) × 50 cm (height) pots during the normal growth season in Huaian, China. Wheat stems at the jointing stage were harvested for protein extraction. Protein was extracted according to Damerval et al. (1986) with modifications. Stems were ground and homogenised, and the proteins were precipitated at –20°C with acetone containing 10% (w/v) trichloroacetic acid, 0.07% (w/v) dithiothreitol, and 150 mM phenylmethyl sulphonyl fluoride for 1 h. After centrifugation for 30 min at 15,000 × *g*, 4°C, the precipitates were washed until colourless with ice-cold acetone containing 0.07% (w/v) dithiothreitol and 150 mM phenylmethyl sulphonyl fluoride to remove polysaccharides, pigments, and lipids, and then dried by vacuum centrifugation. The residue was suspended for 30 min at room temperature in 7 M urea, 2 M thiourea, 4% (w/v) 3-[(3-cholamidopropyl)-dimethylammonio]-1-propane sulphonate, 20 mM dithiothreitol, 0.2% (v/v) carrier ampholyte (pH 3.0–10.0, Amersham Biosciences, Piscataway, NJ, USA), protease inhibitor cocktail (1 µl per 30 mg plant tissue; Sigma, St. Louis, MO, USA). While the precipitate was suspended, it was sonicated five times for 30 s each on ice. The supernatant was obtained by centrifugation at 40,000 × *g* for 30 min and stored at –80°C. The amount of protein in the extract was quantified using the reagents of a 2-D Quant kit (Amersham Biosciences, Piscataway, NJ, USA).

2-DE

Proteins (a 300 g sample) were first separated by isoelectric focusing using immobilised linear pH gradient strips (pH 3.0–10.0, 13 cm) in an Ettan IPGphor 3 apparatus (Amersham Biosciences, Piscataway, NJ, USA). The strips were passively rehydrated for 13 h. The voltage–time settings were: 250 V, 1 h; 500 V, 1 h; 2000 V, 1 h; then 8000 V until 50,000 Vh was reached. The second-dimension separation used SDS-PAGE (12.5% polyacrylamide gels) at

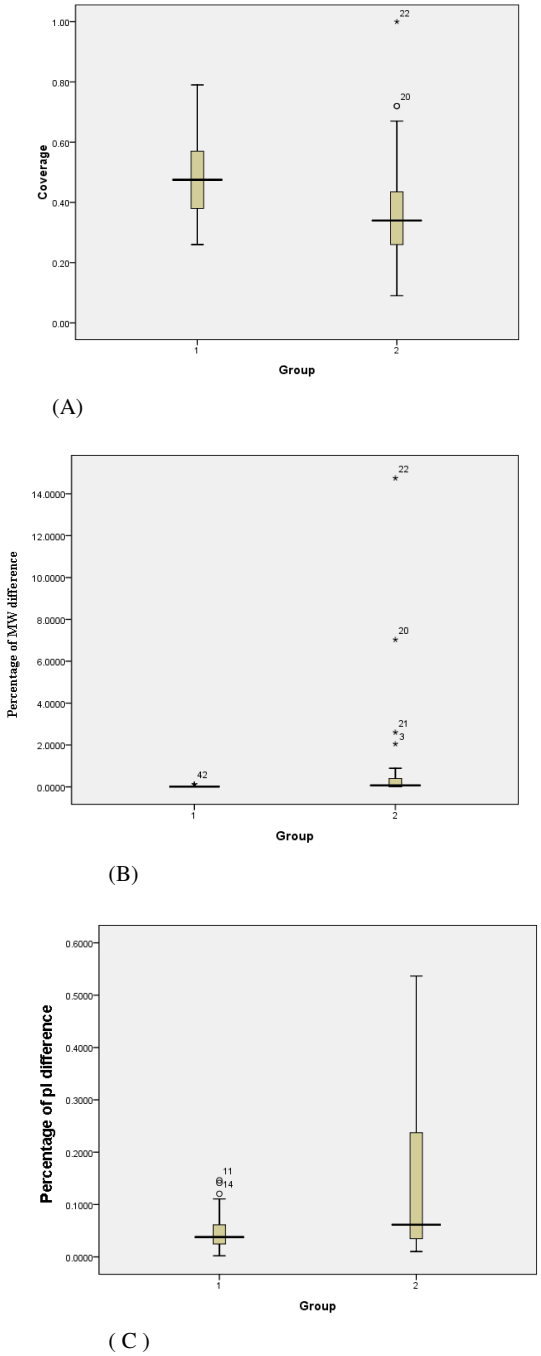

Fig 4. Box-and-whisker plots for sequence coverage, molecular weight differences, and pI differences for the experimental and database-retrieved proteins. A, Sequence coverage for the identified proteins by matching their experimental peptide masses. B, Differences between experimental molecular weights derived from electrophoretic separation of the protein spots and calculated molecular weights of the retrieved proteins. C, Differences between the experimental isoelectric point (pI) values and the calculated pI values. Group 1, Proteins retrieved from the wheat protein/peptide database; Group 2, Proteins retrieved from the green plant NCBInr protein database.

1 W/gel for 90 min and 10 W/gel for 4 h, and a Multiphor II Electrophoresis System (Amersham Biosciences, Piscataway, NJ, USA). Three replicate gels were prepared under the same conditions and were subjected to silver staining to visualise the protein spots (Shevchenko *et al.,* 1996).

MS of tryptic digests of the electrophoretically separated protein spots and protein identification

The stained gel profiles were characterised and quantified using PDQuest software (Bio-Rad, Hercules, CA, USA). Fifty weakly or moderately stained protein spots were cut out of the gel and were each digested with trypsin (Promega, Madison, WI, USA). Mass measurements were acquired in the positive-ion and reflector modes using a MALDI-TOF mass spectrometer (Reflex III, Bruker-Daltonics, Bremen, Germany). Peptide spectra were automatically processed for baseline correction, noise removal, peak deisotoping, and threshold adjustment (2% base peak intensity), according to Watson et al. (2003). External calibration was performed for the mass range 600–3500 Da. Internal calibration was performed using the trypsin auto-digestion fragment residues 108–115 ($[M+H]^+$, 842.509) and residues 58–77 ($[M+H]^+$, 2211.104), which were present in all spectra. The annotations of the peak masses generated by Flexanalysis software, which was provided by Bruker Daltonics, were manually checked and edited to ensure that the monoisotopic peaks had been correctly labelled. The edited peak lists were used for the database searches. Mass fingerprinting searches were performed using a local Mascot program (http://www.matrixscience.com) against the wheat database that had been constructed from the wheat EST database (from the NCBI dbEST database, http://ftp.ncbi.nlm.nih.org)**,** as described in Results**,** and the NCBInr green plant protein database with the following parameters: monoisotopic peptide mass ($[M+H]^+$), one allowed incomplete trypsin cleavage per peptide, a mass tolerance of 50 ppm, a requirement for carboxyamidomethyl cysteines, and allowance for methionine oxidation. A protein was considered to be identified when the following criteria were met: at least four matching peptides and >20% sequence coverage. Electrospray ionisation-tandem MS analysis of the tryptic peptides from protein spots and protein identification were performed accorded to Yang et al. (2010).

Acknowledgements

This project was supported by grants from National Science Foundation of China (No.30901156 and No.31170619 to Jinhui Chen; No.30930077 to Jisen Shi; No.30871704, No.30971452 and No.31170256 to Xiangyang Hu) and Chinese National Forest Bureau's '948' grant (No. 2009-4-24) and Natural Science Foundation of Jiangsu University (No. 09KJA220001) to Jinhui Chen.

References

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

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Nucleic Acids Res 25:3389-3402

Damerval C, Devienne D, Zivy M, Thiellement A (1986) Technical improvements in two-dimensional electrophoresis increase the level of genetic variation detected in wheat-seedling proteins. Electrophoresis 7:52-54

Dong GJ, Pan WD, Liu GS (2006) The analysis of proteome changes in sunflower seeds induced by N^+ implantation. J Biosciences 31:247-253

Edwards NJ (2007) Novel peptide identification from tandem mass spectra using ESTs and sequence database compression. Mol Syst Biol 3:102

Gracheva EO, Ingolia NT, Kelly YM, Cordero-Morales JF, Hollopeter G, Chesler AT, Sánchez EE, Perez JC, Weissman JS, Julius D (2010) Molecular basis of infrared detection by snakes. Nature 464:1006-1011

Grimplet J, Gasper JW, Gancel A, Sauvage F, Romieu C (2005) Including mutations from conceptually translated expressed sequence tags into orthologous proteins improves the preliminary assignment of peptide mass fingerprints on non-model genomes. Proteomics 5:2769-2777

Habermann B, Oegema J, Sunyaev S, Shevchenko A (2004) The power and the limitations of cross-species protein identification by mass spectrometry-driven sequence similarity searches. Mol Cell Proteomics 3:238-249

Higashi Y, Hirai MY, Fujiwara T, Naito S, Noji M, Saito K (2006) Proteomic and transcriptomic analysis of Arabidopsis seeds: molecular evidence for successive processing of seed proteins and its implication in the stress response to sulfur nutrition. Plant J 48:557-571

Huang M, Chen T, Chan Z (2006) An evaluation for cross-species proteomics research by publicly available expressed sequence tag database search using tandem mass spectral data. Rapid Commun Mass Sp 20:2635-2640

Huang X, Madan A (1999) CAP3: A DNA sequence assembly program. Genome Res 9:868-877

Irar S, Brini F, Goday A, Masmoudi K, Pagès M (2010) Proteomic analysis of wheat embryos with 2-DE and liquid-phase chromatography (ProteomeLab PF-2D) - A wider perspective of the proteome. J Proteomics 73:1707-1721

Kamal AHM, Kim KH, Shin DH, Seo HS, Shin KH, Park CS, Heo HY, Woo SH (2009) Profile of pre-harvest sprouting wheat by using MALDI-TOF Mass Spectrometry. Plant Omics J. 2:110-119

Kamal AHM, Kim K-H, Shin K-H, Choi J-S, Baik B-K, Tsujimoto H, Heo HY, Park C-S, Woo S-H (2010) Abiotic stress responsive proteins of wheat grain determined using proteomics technique. Aust J Crop Sci 4: 196-208

Kim SI, Kim JY, Kim EA, Kwon KH, Kim KW, Cho K, Lee JH, Nam MH, Yang DC, Yoo JS, Park YM (2003) Proteome analysis of hairy root from *Panax Ginseng C.A. Meyer* using peptide fingerprinting, internal sequencing and expressed sequence tag data. Proteomics 3:2379-2392

Kwon KH, Kim M, Kim JY, Kim KW, Kim SI, Park YM, Yoo JS (2003) Efficiency improvement of peptide identification for an organism without complete genome sequence, using expressed sequence tag database and tandem mass spectral data. Proteomics 3: 2305-2309

Laino P, Shelton D, Finnie C, De Leonardis AM, Mastrangelo AM, Svensson B, Lafiandra D, Masci S (2010) Comparative proteome analysis of metabolic proteins from seeds of durum wheat (cv. Svevo) subjected to heat stress. Proteomics 10:2359-2368

Lisacek FC, Traini MD, Sexton D, Harry JL, Wilkins MR (2001) Strategy for protein isoform identification from expressed sequence tags and its application to peptide mass fingerprinting. Proteomics 1:186-193

Mallick P, Kuster B (2010) Proteomics: a pragmatic perspective. Nat Biotechnol 28:695-709

Mathesius U, Imin N, Chen HC, Djordjevic MA, Weinman JJ, Natera SHA, Morris AC, Kerim T, Paul S, Menzel C, Weiler GF, Rolfe BG (2002) Evaluation of proteome reference maps for cross-species identification of proteins by peptide mass fingerprinting. Proteomics 2:1288-1303

Mathesius U, Keijzer G, Natera SHA, Weinman JJ, Djordjevic MA, Rolfe BG (2001) Establishment of a root proteome reference map for the model legume Medicago truncatula using the expressed sequence tag database for peptide mass fingerprinting. Proteomics 1: 1424-1440

Merlino M, Leroy P, Chambon C, Branlard G (2009) Mapping and proteomic analysis of albumin and globulin proteins in hexaploid wheat kernels (*Triticum aestivum* L.). Theor Appl Genet 118:1321-1337

Mooney BP, Krishnan HB, Thelen JJ (2004) High throughput mass fingerprinting of soybean seed proteins: Automated workflow and utility of unigene expressed sequence tag databases for protein identification. Phytochemistry 65:1733-1744

Neale DB, Kremer A (2011) Forest tree genomics: growing resources and applications. Nat Rev Genet 12:111-122

Østergaard O, Finnie C, Laugesen S, Roepstorff P, Svennson B (2004) Proteome analysis of barley seeds: Identification of major proteins from two-dimensional gels (pI 4-7). Proteomics 4:2437-2447

Palagi PM, Lisacek F, Appel RD (2009) Database interrogation algorithms for identification of proteins in proteomic separations. Methods Mol Biol 519:515-531

Peng Z, Wang M, Li F, Lv H, Li C, Xia G (2009) A proteomic study of the response to salinity and drought stress in an introgression strain of bread wheat. Mol Cell Proteomics 8: 2676-2686

Porubleva L, Vander Velden K, Kothari S, Oliver D J, Chitnis PR (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

Rajjou L, Belghazi M, Huguet R, Robin C, Moreau A, Job C, Job D (2006) Proteomic investigation of the effect of salicylic acid on Arabidopsis seed germination and establishment of early defense mechanisms. Plant Physiol 141:910-923

Rice P, Longden I, Bleasby A (2000) EMBOSS: The European Molecular Biology Open Software Suite. Trends Genet 16:276-277

Scippa GS, Rocco M, Ialicicco M, Trupiano D, Viscosi V, Di Michele M,, Arena S, Chiatante D, Scaloni A (2010) The proteome of lentil (Lens culinaris Medik.) seeds: discriminating between landraces. Electrophoresis 31:497-506

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

Shin DH, Kamal AHM, Suzuki T, Yun UH, Lee MS, Chuang KY, Jeong HS, Park CH, Choi JS, Woo SH (2010) Reference proteome map of buckwheat (Fagopyrum esculentum and Fagopyrum tataricum) leaf and stem cultured under light or dark. Aust J Crop Sci 4:633-641

Watson BS, Asirvatham VS, Wang L, Sumner LW (2003) Mapping the proteome of barrel medic (Medicago truncatula). Plant Physiol 131:1104-1123

Wright JC, Beynon RJ, Hubbard SJ (2010) Cross species proteomics. Methods Mol Biol 604:123-135

Yahata E, Maruyama-Funatsuki W, Nishio Z, Tabiki T, Takata K, Yamamoto Y, Tanida M, Saruyama H (2005) Wheat cultivar-specific proteins in grain revealed by 2-DE and their application to cultivar identification of flour. Proteomics 5:3942-3953

Yang LM, Luo YM, Wei JF, Ren CM, Zhou X, He SH (2010) Methods for protein identification using expressed sequence tags and peptide mass fingerprinting for seed crops without complete genome sequences. Seed Sci Res 20:257-262

Yang P, Li X, Wang X, Chen H, Chen F, Shen S (2007) Proteomic analysis of rice (*Oryza sativa*) seeds during germination. Proteomics 7:3358-3368

Yates JR (2000) Mass spectrometry. From genomics to proteomics. Trends Genet 16:5-8