# Current status of soybean proteomics

**Contents**



## 1. Abstract

Although research on non-legume model species such as *Arabidopsis thaliana* and rice provides insight into many fundamental aspects of plant biology, it cannot address some important aspects of legume biology. Legumes are of immense importance to human and an important crop for sustainable agriculture. Two model species, *Lotus japonicas* and *Medicago truncatula*, would have been the focus of genome sequencing and functional genomics programmes. Unfortunately, agricultural legumes are relatively poor model systems for genetics and genomics research. Even though soybean is an important crop to supply a major portion of the world's demand for vegetable oil and protein, the sequencing of the soybean genome is in its infancy. So, proteomics would be a powerful tool for its functional analysis.

## 2. Introduction

Food shortage is the most serious problem of current century worldwide. To meet the expanding food demands of the rapidly growing world population, crop production will need to increase by a further 50% by 2025 (Khush, 2003). Rice, wheat and maize provide approximately half of the calories consumed by the world's population. Furthermore, soybean and other oilseeds are significant source of fatty acids and proteins for human and animal nutrition as well as for nonedible uses, including industrial feedstocks and combustible fuel (Thelen and Ohlrogge, 2002). However, almost all the cultivated land is under sub-optimal conditions for plant growth. About 70% of yield potential has been estimated to be lost by unfavorable physiochemical environments, even in developed agricultures (Boyer, 1982). To meet these challenges, genes and proteins that control crop plants architecture and/or stress resistance in a wide range of environments need to be identified and characterized to facilitate the molecular improvement of crop productivity.

Although research on non-legume model species such as *Arabidopsis thaliana* (*A. thaliana*) (The Arabidopsis Genome Initiative, 2000) and rice (International Rice Genome Sequence Project, 2005) provided insight into many fundamental aspects of plant biology, it cannot address some important aspects of legume biology. Legumes are of immense importance to humanity and a key in sustainable agriculture. Two model species, *Lotus japonicas* and *Medicago truncatula* (*M. truncatula*), are the focus of genome sequencing and functional genomics (Udvardi et al., 2005). Unfortunately, agricultural legumes such as soybean are relatively poor model system for genetics and genomics research, because soybean has genome duplications and self-incompatible or have a long generation time/ Although soybeans are important crop to supply a major portion of the world’s demand for vegetable oil and protein, the sequencing of the

soybean genome is in its infancy. In this case, proteomics approach will be a powerful tool for analysis of plant macromolecules function.

Gaining an understanding of the biological function of any novel gene is a more ambitious goal than just obtaining their sequences. The wealth of information on nucleotide sequences being generated through genome projects far outweighs what is currently available on amino acid sequences of known proteins (Lockhart and Winzeler, 2000; Pandey and Mann, 2000). The information with genome sequence data and inferred protein sequence data can be used to identify proteins and to follow changes in protein expression through time-dependency in an organism. Recently, *M. truncatila* has been the subject of several proteomic studies. The proteome of *M. truncatila* cell suspension culture was analyzed using two-dimensional polyaclylamide gel electrophoresis (2D-PAGE) and nanoscale high performance liquid chromatography (HPLC) coupled to a tandem quadro type time-of-flight mass spectrometer (Q-TOF MS) to yield an extensive protein reference map (Lei et al., 2005). Furthermore, a proteome study of *M. truncatila* protoplast cultures was done to investigate molecular changes taking place during protoplast proliferation (de Jong et al., 2007). These works represent the most extensive proteomic description of *M. truncatula* suspension cells to date and provide a reference map for future comparative proteomics and functional genomics studies of biotic and abiotic stress responses.

In spite of the soybean's importance in agriculture, yield increment of this crop through conventional breeding over the past few decades have lagged behind those of cereals. Numerous abiotic and biotic impediments including flooding, drought, salinity, acidity and nutrient limitations, and various diseases and pests curtail yield potential in soybean. Although research on *M. truncatila* provides insight into some

fundamental aspects of molecular biology, it cannot address any important aspects of food legumes such as soybean. The purpose of this review is to discuss the strength and weakness of proteomics approach and limitations of current techniques for soybean molecular biology.

**3. Comprehensive proteomics analysis of soybean**

The knowledge of the cellular proteome facilitates the identification of changes in protein expression under different growing condition and treatments. The analytical methodology for the separation and identification of large numbers of proteins should be reliable and verifiable, and meet following criteria: (1) It should possess an excellent analytical resolution in the separation of proteins and high accuracy in protein identification, (2) it should resolve the extensive molecular variety of proteins resulted from alternative splicing, mRNA editing, or co- and post-translational modifications, and (3) it should allow for a methodological performance in a high-throughput fashion. Several attempts have been made to create proteome map of soybean using 2D-PAGE.

Mooney et al. (2004) adapted peptide mass fingerprinting to identify some of the previously detected proteins on 2D-PAGE and facilitate high-throughput sample processing. They outline a workflow employing robotic automation at each step subsequent to 2D-PAGE and demonstrate the utility of UniGene databases in proteomic investigations. Total protein from mature dry soybean (*Glycine max* cv. Jefferson) seed was isolated and 2D-PAGE performed using 13 cm Immobilized pH gradient (IPG) strips and Hoeffer SE600 units for isoelectric focusing and SDS–PAGE, respectively. Protein spots were analyzed using Phoretix 2D-Advanced software. Excised protein spots arrayed into 96-well plates and transferred to a Multiprobe II EX liquid handling

station for subsequent destaining, tryptic digestion and peptide extraction. The matrix-assisted laser desorption ionization (MALDI) -TOF MS was operated in the positive ion delayed extraction reflector mode. Peptide spectra were submitted to a MS Fit program of Protein Prospector. Assignments from UniGene contigs were subsequently searched against the NCBI non-redundant database using the BLASTP search algorithm to determine similarity matches. They have found that insertion of a phenol extraction step immediately after seed crushing resulted in less contamination with polysaccharides, nucleic acids or other non-protein macromolecules. Ninety six out of 128 protein spots detected by Coomassie Brilliant Blue (CBB) stained 2D-PAGE were subjected to automated peptide mass fingerprinting. Because of broad dynamic range in protein expression, the author suggested for pre-fractionation steps or narrow range IPG strips which ultimately allow for greater protein loads to increase the number of detectable spots. Seventeen of 44 proteins identified were assigned to β-conglycinin subunits and glycinin which belongs to two major storage proteins 7S globulins and 11S globulins already characterized in soybean. Some abundant, non seed storage proteins were also found such as six different sucrose binding proteins, one alcohol dehydrogenase and two seed maturation proteins. They suggested that proteome surway could help seed physiologist to determine the function of unknown proteins.

Since protein extraction methods for soybean seeds have not been studied in details as other techniques, Natarajan et al (2005) have compared four different protein extraction/solubilization methods, which are urea, thiourea/urea, phenol, and a modified trichloroacetic acid (TCA)/acetone, for accurate analysis of soybean (cv. Williams 82) seed proteins through 2D–PAGE. Urea and phenol methods resolved fewer protein spots whereas thiourea/urea and TCA/acetone methods were efficient and reliable for

2D-PAGE separation of soybean seed proteins. The TCA/acetone and thiourea/urea methods removed nonprotein and proteolytic components that interfere with isoelectric focusing (IEF). A total of 15 spots recovered from the modified TCA/acetone method and subjected to the MALDI–TOF MS and LC–MS led to good quality spectra, indicating the compatibility of the TCA/acetone method with MS analysis. Two major storage proteins of soybean seeds, β-conglycinin, and both acidic and basic glycinin polypeptide chains, were well separated using all four extraction procedures. Less abundant nonstorage proteins such as alcohol dehydrogenase, trypsin inhibitor, allergen protein and sucrose binding protein precursors were either faint or non-detectable using the urea and phenol methods but was clearly resolved using the TCA/acetone and thiourea/urea methods.

Despite the importance of seed filling in the synthesis of storage reserves for germination, systematic proteomic analysis of this phase in legumes is yet to be carried out. Hajduch et al. (2005) analyzed soybean (cv. Maverick) seed proteins at 14, 21, 28, 35 and 42 days after flowering using 2D-PAGE. They started with IPG strips of pH 3 to 10, but narrow down the IEF pH range to 4 to 7 for high-resolution proteome maps. Two normalizations were performed to compensate the differences in number and abundances of protein spots. A total of 488 and 191 proteins were identified from 2D-PAGE gels of pH range 4 to 7and 3 to 10 gels (pH 7–10 region only), respectively. Each of the 679 proteins was excised from reference gels for identification by MALDI-TOF MS and a total of 422 proteins (62%) were identified. One unique protein was often represented by more than one spot on the 2D-PAGE gel, most likely due to post-translation modifications or genetic isoforms. Taking into account this redundancy, 216 unique proteins out of 422 were identified. A total of 82 proteins

associated with metabolism, which was the largest functional class. The second largest functional class was comprised of 52 spots assigned to the seed storage proteins β-conglycinin and glycinin. They suggested that genetic redundancy was one of the possible explanation for the multiple species observed within these proteins. An overall down- and up-regulation was observed for metabolism and storage related proteins, respectively, during seed filling, suggesting metabolic activity curtails as seeds approach maturity. Abundance of proteins related to metabolite transporter, disease and defense, energy production, cell growth and division, signal transduction, protein synthesis and secondary metabolism did not vary significantly Based upon the similarities between expression profiles of 92 unknown proteins and those of the major functional class, they were able to classify them into five expression profile groups. Thirteen sucrose-binding proteins mapped to the same UniGene accession number reported earlier and shown to be involved in sucrose transport. This result support previous investigations on the importance of sucrose as a signaling molecule in seed and embryo development.

To investigate the molecular mechanism of cadmium detoxification, Sobkowiak and Deckert (2006) have treated suspension culture of soybean (cv. Navik) cells with various concentrations of 3, 5, 6 and 10 μM $Cd^{2+}$ and labeled with [$^{35}$S]-methionine for 24, 48 and 72 h. The protein accumulation was analyzed by SDS-PAGE following CBB staining, whereas the synthesis of [$^{35}$S]-labeled protein was detected by autoradiography. The result showed preferential accumulation of polypeptides of molecular weight 16 kDa and the synthesis of polypeptides of 26 and 41 kDa which was dose- and time-dependent. These bands were analyzed through electrospray ionization (ESI) Q-TOF MS and the protein fragments were identified by searching protein

database using Mascot. The 26-kDa polypeptide band contained superoxide dismutase (Cu–Zn) and histone H2B which was not observed in control. The former is known constituents of plant defense reaction against metal toxicity including soybean cells treated with $Cd^{2+}$and the later might be required for repair of $Cd^{2+}$-induced DNA damage. The 26-kDa polypeptide band contained various glutathione S-transferases which have been reported in detoxification of wide range of xenobiotic compounds, including heavy metals. The third polypeptide band showed homology to chalcone synthase which is activated in plants under various stress conditions.

The connection between function of some soybean peribacteroid membrane proteins and the primary structure of others has been studied and finally isolated genes encoding novel peribacteroid membrane proteins for functional studies (Panter et al., 2000). Panter et al. (2000) have purified the peribacteroid membrane from isolated symbiosomes and using cytochrome C reductase activity as a marker for the bacteroid inner membrane demonstrated that the peribacteroid membrane contamination by other membranes was less than 1% on a protein basis. Proteins from purified peribacteroid membrane separated by 2D-PAGE in which CBB-stained gels were less sensitive than silver staining of similar gels in detection of proteins. Thirty-one out of 100 protein spots identified by CBB stained gels were subjected to N-terminal sequencing reactions. Comparing N-terminal sequences with protein sequences present in various public data bases, using Fasta3, showed six putative peribacteroid membrane proteins homologous to those of known proteins involved in translocation, processing or degradation of other proteins. The sequences of remaining two proteins out of eight identified were identical to each other and to the sequence of heat shock protein (HSP) 60 from maize mitochondria. Presence of chaperones homolog such as HSP60 and binding protein

from the HSP70 suggested that symbiosomes may import some proteins directly from the cytoplasm in vivo. The pathway of protein import from the cytoplasm may allow the cell greater flexibility in tailoring the protein complement of symbiosomes. Nine sequences did not show significant similarity to any other proteins. All of the putative peribacteroid membrane proteins that matched known proteins were peripheral rather than integral membrane proteins because of inefficient solubilization and recovery of the hydrophobic proteins during sample preparation, aggregation of hydrophobic proteins during IEF and poorly staining by CBB as it preferentially stains arginine and lysine.

**4. Proteome databases of soybean**

The genomes of rice and *A. thaliana* have been sequenced. Powerful genomic approaches for these plants now onward include microarrays to examine changes in transcript levels and knockout lines for most of the genes. Because proteins are the master keys in most processes of living cells, many plant proteomic investigations have been reported recently, and several plant proteome databases have also been constructed. The *A. thaliana* proteome databases were more extensive than other plant proteome databases, because *A. thaliana* genome has been sequenced completely and deduced protein sequences could be retrieved from the EMBL proteome site (http://www.ebi.ac.uk/proteome). In addition to the EMBL resources, many other proteome data sets for *A. thaliana* have also been accumulating (Peck, 2005). Rice Proteome Database based on 2D-PAGE is also available on the web site (http://gene64.dna.affrc.go.jp/RPD/main_en.html) (Komatsu et al., 2004). Proteins extracted from rice tissues, some cellular organelles of rice, and rice under various kinds

of stresses were separated by 2D-PAGE. It was revealed using an image analyzer that protein spots could be detected on 2D-PAGE gels stained by CBB, and determined using the protein sequencer or MALDI-TOF MS. Finally, Rice Proteome Database was constructed which included information on amino acid sequences and sequence homologies. The protein sequences of other plants were then retrieved from the Swiss-Prot (release 40) and TrEMBL (release 20) data banks (Boeckmann et al., 2003). 2D-PAGE gel protein reference maps of sub-proteomes of different plant species are expected to become a central tool for organizing and understanding the plant proteome. Web sites with organized 2D-PAGE databases are already available (http://sphin.rug.ac.be:8080/ppmdb/index.html, http://www.biokemi.su.se/chloroplast/ and http://www.expasy.ch/ch2D/). In the future, reference 2D-PAGE maps will be used to follow differential protein expression and post-translational modifications.

A high-throughput proteomic approach was employed to determine the expression profile and identity of 100 proteins during seed filling in soybean (cv. Maverick). Soybean seed proteins were analyzed at 2, 3, 4, 5 and 6 weeks after flowering using 2D-PAGE and MALDI-TOF MS. A user-intuitive database (http://oilseedproteomics.missouri.edu) was developed to access these data for soybean and other oilseeds currently being investigated. This led to the establishment of high-resolution proteome reference map, expression profiles of 679 spots, and corresponding MALDI-TOF MS spectra for each spot (Hajduch et al., 2005).

Soybean Proteome Database has been also constructed by Komatsu et al. (in this database). A high-throughput proteomic approach was employed to determine the expression profile and identity of 250 proteins of soybean (cv. Enrei) development. Soybean proteins from embryonic axic, cotyledon, root, hypocotyl and so on were

separated by 2D-PAGE, stained by CBB, and analyzed by protein sequencer and MS. Finally, Soybean Proteome Database was constructed which included information on amino acid sequences and sequence homologues.

**5. Differential proteomics analysis of soybean**

In recent years, the application of proteomic tools such as 2D-PAGE, MALDI-TOF MS and LC-MS/MS has become popular, and these tools are powerful methodologies for accurately detecting and examining changes in protein composition. These tools have been extensively used to examine the composition of protein profiles of both natural and transgenic plant, control and treatment, and/or wild type and mutant. However, limited studies are available for detecting abundant and nonabundant proteins at the subunit level, because it remains a challenging issue (Herman et al., 2003).

Symbiotic interactions between legume plants and rhizobia induce specific metabolisms and intracellular organelles in nodules. To survey symbiotic differentiation of key organelle, mitochondria, protein constituents of nodule and root mitochondria in soybean were compared after 2D-PAGE, and the proteins were characterized in combination with MALDI-TOF MS and protein sequencer (Hoa et al., 2004). Of the proteins that were only detected in nodule mitochondria, phosphoserine aminotransferase, flavanone 3-hydroxylase, coproporphyrinogen III oxidase, one ribonucleoprotein and three unknown proteins were identified. Seven up-regulated and 8 down-regulated protein spots in nodule mitochondria were also assigned protein identities. The physiological roles of these differential expressions were discussed in relation to nodules-specific metabolism in soybean nodules.

Infection of *Bradyrhizobium japonicum* (*B. japonicum*) to soybean root hairs is

the first among several complex events leading to nodulation. Wan et al. (2005) reported about the proteins of soybean root hairs after inoculation with *B. japonicum*. Following protein separation by 2D-PAGE, in one experiment, 96 protein spots were analyzed by MALDI-TOF MS to compare protein profiles between uninoculated roots and root hairs. Another 37 spots, derived from inoculated root hairs over different time points, were also analyzed by tandem MS (MS/MS). As expected, some proteins were differentially expressed in root hairs compared with roots (e.g., a chitinase and phosphoenolpyruvate carboxylase). Out of 37 spots analyzed by MS/MS, 27 candidate proteins were identified by database comparisons. These included several proteins known to respond to rhizobial inoculation (e.g., peroxidase and phenylalanine-ammonia lyase). However, novel proteins were also identified (e.g., phospholipase D and phosphoglucomutase). This research established an excellent system for the study of root-hair infection by rhizobia and in a more general sense, the functional genomics of a single plant cell type. The results also indicated that proteomic studies in soybean, lacking a complete genome sequence, are worthwhile.

Natarajan et al. (2006) reported characterization of storage proteins in wild (*Glycine soja*) and cultivated (*Glycine max*) soybean seeds using proteomic analysis. A combined proteomic approach was applied for the separation, identification, and comparison of two major proteins, β-conglycinin and glycinin, in wild and cultivated soybean seeds. 2D-PAGE with three different IPG strips was an effective method to separate a large number of abundant and less-abundant storage proteins. Most of the β-conglycinin subunits were well-separated in the pH range 3 – 10, while acidic and glycinin polypeptides were well-separated in pH range 4 – 8 and 6 – 11, respectively. Although the overall distribution pattern of the protein spots was similar in both

genotypes using pH 3 – 10, variations in number and intensity of protein spots were better resolved using a combination of pH 4 – 7 and pH 6 – 11. The total number of storage protein spots detected in wild and cultivated genotypes was 44 and 34, respectively. They have demonstrated that high-resolution near-infrared has a potential to identify heat-stable antinutrirional factors, such as nonstarch polysaccharide and oligosaccharides in soybean. This can aid in the selection of suitable soybean for use in diets for target monogastric species.

## 6. Concluding remarks

Although current depth of coverage for soybean proteome is still significantly less than other plants, the reference map of soybean provides a starting point for ongoing functional genomics studies associated with biotic/abiotic stress and natural product biosynthesis in soybean. Continued research towards the development of a soybean proteome map will be useful for rapid comparison of soybean cultivars, mutants and transgenic lines. While investigations into compositional analysis, soybean physiology will also benefit from a detailed and quantitative proteome reference map of soybean plant. The information obtained from soybean proteomics will be helpful in predicting the function of plant proteins and aid in molecular cloning of corresponding genes in near future. Identification of novel genes, determination of their expression patterns in response to stress, and understanding of their functions to stress adaptation will provide us the basis for effective engineering strategies to improve soybean stress tolerance.

## 7. References

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.-C., Estreicher, A., Gasteiger, E.,

Martin, M.J., Michoud, K., O'Donovan, C., Phan, I., et al. (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucl Acids Res* **31**: 365-70.

Boyer, J. S. (1982). Plant productivity and environment. *Science* **218:** 443–448.

Hajduch, M., Ganapathy, A., Stein, J.W. and Thelen, J.J. (2005). A systematic proteomic study of seed filling in soybean. Establishment of high-resolution two-dimensional reference maps, expression profiles, and an interactive proteome database. *Plant Physiol.* **137**: 1397-419.

Herman, E. M., Helm, R. M., Jung, R. and Kinney, A. J. (2003). Genetic modification removes an immunodominant allergen from soybean. *Plant Physiol.* **132**: 36-43.

Hoa le TP, Nomura M, Kajiwara H, Day DA, Tajima S. (2004). Proteomic analysis on symbiotic differentiation of mitochondria in soybean nodules. Plant Cell Physiol. 2004, 45(3):300-8.

International Rice Genome Sequencing Project (2005). The map-based sequence of the rice genome. *Nature* **436**: 793-800.

Khush, G.S. (2003). Challenges for meeting the global food and nutrient needs in the new millennium. *Proc. Nutr. Soc.* **60**: 15-26.

Komatsu S, Kojima K, Suzuki K, Ozaki K, Higo K. (2004). Rice Proteome Database based on two-dimensional polyacrylamide gel electrophoresis: its status in 2003. *Nucl Acids Res* 32: 388-392.

Lei, Z., Elmer, A.M., Watson, B.S., Dixon, R.A., Mendes, P.J. and Sumner, L.W. (2005). A two-dimensional electrophoresis proteomic reference map and systematic identification of 1367 protein from a cell suspension culture of the model legume Medicago truncatula. *Mol. Cell. Proteomics* **4**: 1812-25.

Lockhart, J.D. and Winzeler, A.E. (2000). Genomics, gene expression and DNA arrays. *Nature* **405**: 827-35.

Mooney, B.P., Krishnan, H.B. and Thelen, J.J. (2004). High-throughput peptide mass fingerprinting of soybean seed proteins: automated workflow and utility of UniGene expressed sequence tag databases for protein identification. *Phytochemistry* **65**: 1733-44.

Natarajan, S., Xu, C., Caperna, T.J. and Garrett, W.M. (2005). Comparison of protein solubilization methods suitable for proteomic analysis of soybean seed proteins. *Anal. Biochem.* **342**: 214-20.

Natarajan, S.S., Xu, C., Bae, H., Caperna, T.J. and Garrett, W.M. (2006). Characterization of storage proteins in wild (Glycine soja) and cultivated (Glycine max) soybean seeds using proteomic analysis. *J Agric Food Chem.* **54**: 3114-20.

Pandey, A. and Mann, M. (2000). Proteomics to study genes and genomics. *Nature* **405**: 837-45.

Panter, S., Thomson, R., Bruxelles, G. de, Laver, D., Trevaskis, B. and Udvardi, M. (2000). Identification with proteomics of novel proteins associated with the peribacteroid membrane of soybean root nodules. *Mol. Plant-Microbe Interact.* **13**: 325–333.

Peck, S.C. (2005). Update on proteomics in Arabidopsis. Where do we go from here? *Plant Physiol* **138**: 591-9.

Sobkowiak, R. and Deckert, J. (2006). Proteins induced by cadmium in soybean cells. *J. Plant Physiol.* **163**: 1203-1206.

The Arabidopsis Genome Initiative (2000). Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* **408**: 796-815.

Thelen, J. and Ohlrogge, J. (2002). Metabolic engineering of fatty acid biosynthesis in plants. *Metabolic Engineering* **4**: 12-21.

Udvardi, M.K., Tabata, S., Parniske, M. and Stougaard, J. (2005). Lotus japonicus: legume research in the fast lane. *Trends Plant Sci.* **10**: 222-8.

de Jong, F., Mathesius, U., Imin, N. and Rolfe, B.G. (2007). A proteome study of the proliferation of cultured Medicago truncatula protoplasts. *Proteomics* **7**: 722-36.

Wan, J., Torres, M., Ganapathy, A., Thelen, J., DaGue, B.B., Mooney, B., Xu, D. and Stacey, G. (2005). Proteomic analysis of soybean root hairs after infection by Bradyrhizobium japonicum. *Mol Plant Microbe Interact.* **18**: 458-67.