BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Database

Volume 10(9)

SoyProLow: A protein database enriched in low abundant soybean proteins

Mona Tavakolan[1], Nadim W. Alkharouf[1], Benjamin F. Matthews[2] & Savithiry S. Natarajan[2]*

[1]Department of Computer and Information Sciences, Towson University, Towson, MD 21252, USA; [2]USDA-ARS, Soybean Genomics and Improvement Laboratory, Beltsville, MD 20705, USA; Savithiry S. Natarajan - Email: savi.natarajan@ars.usda.gov; Phone Number: 301-504-5258; Fax Number: 301-504-5728; *Corresponding author

Received August 28, 2014; Accepted August 31, 2014; Published September 30, 2014

Abstract:
Soybeans are an important legume crop that contain 2 major storage proteins, β-conglycinin and glycinin, which account about 70-80% of total seed proteins. These abundant proteins hinder the isolation and characterization of several low abundant proteins in soybean seeds. Several protein extraction methodologies were developed in our laboratory to decrease these abundant storage proteins in seed extracts and to also decrease the amount of ribulose-1, 5-bisphosphate carboxylase/oxygenase (RuBisCO), which is normally very abundant in leaf extracts. One of the extraction methodologies used 40% isopropanol and was more effective in depleting soybean storage proteins and enhancing low abundant seed proteins than similar methods using 10-80% isopropanol. Extractions performed with 40% isopropanol decreased the amount of storage proteins and revealed 107 low abundant proteins when using the combined approaches of two-dimensional polyacrylamide gel electrophoresis (2D-PAGE) and Mass Spectrometry (MS). The separation of proteins was achieved by iso-electric focusing (IEF) and 2D-PAGE. The proteins were analyzed with MS techniques to provide amino acid sequence. The proteins were identified by comparing their amino acid sequences with those in different databases including NCBI-non redundant, UniprotKB and MSDB databases. In this investigation, previously published results on low abundant soybean seed proteins were used to create an online database (SoyProLow) to provide a data repository that can be used as a reference to identify and characterize low abundance proteins. This database is freely accessible to individuals using similar techniques and can be for the subsequent genetic manipulation to produce value added soybean traits. An intuitive user interface based on dynamic HTML enables users to browse the network and the profiles of the low abundant proteins.

Availability: http://bioinformatics.towson.edu/Soybean_low_abundance_proteins_2D_Gel_DB/Gel1.aspx

Background:
Soybean (*Glycine max*) seeds provide an inexpensive source of protein for both humans and animals and are also a dominant oilseed compared to other legumes. Soybean seeds contain numerous proteins that provide nutrients to the embryo and protect the seed from diseases and pests. Soybean seeds contain two major storage proteins, β -conglycinin and glycinin. These two compounds account for 80 % of total proteins and are responsible for much of the nutritional, physicochemical and physiological properties of soybean seeds. Soybean seeds also contain lower abundance proteins that either provide nutrition or function in metabolism, i.e., β-amylase, cytochrome c, and urease. Other lower abundance proteins may provide defense against diseases and pest. These protective proteins include lectin, lipoxygenases and the Bowman-Birk Inhibitors (BBI) of chymotrypsin and trypsin **[1, 2].** Proteomic analysis using 2D-PAGE for protein separation and MS for protein identification are efficient tools to study alterations in protein profiles that are caused by mutations, the introduction of silencing genes, or to responses to various environmental stimuli **[3, 4].** However, protein extraction for IEF is difficult because seed samples have lipids and other compounds that disrupt or interfere with focusing during protein separation steps. Additionally, it is difficult to detect,

identify, and analyze low abundant proteins since they are obscured by high abundant storage proteins in most legume seeds. Characterization of low abundant soybean seed proteins and the availability of databases for those proteins are limited. Wang *et al.* **[5]** reported that a combination of TCA and phenol was efficient for extracting leaf proteins from bamboo, lemon, olive, redwood. In addition, these authors reported that the TCA method was also suitable for extracting proteins from apple, pear, banana, grape, tomato and orange fruits. In our laboratory, we used direct precipitation of protein from powdered tissues using TCA/acetone to successfully extract proteins from soybean seeds, and leaves **[6, 7].** Using this extraction method, we observed an increased yield of highly abundant proteins with decreased yields of low abundant proteins. Therefore, we developed an efficient isopropanol extraction method that removed approximately 80% of the major seed storage proteins and this allowed us to identify low abundant soybean seed proteins **[8].** We used protein expression profiling to categorize various low abundance proteins. This involved the combined applications of separation techniques such as 2D-PAGE, with identification techniques such as Matrix Assisted Laser Desorption Ionization (MALDI-TOF-MS), Liquid Mass Spectrometry (LC-MS) analysis, and bioinformatics tools.

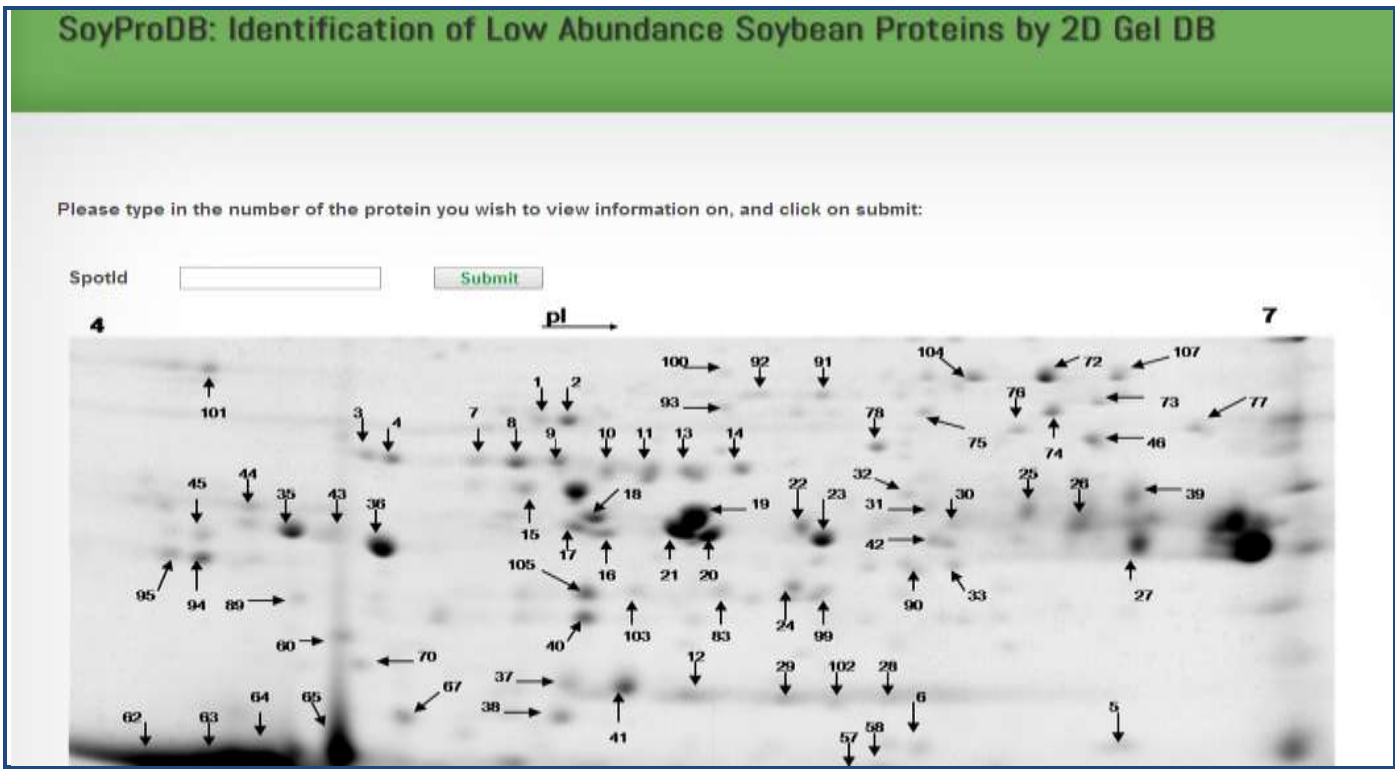

Figure 1: A snapshot from the SoyProLow online database displaying the search field. The numbers on the image represents proteins that have been separated by 2D-PAGE. When the user types the number ID of the protein of interest, the database returns the detailed annotation (if available) for that protein along with other relevant information such as the PI (isoelectric point).

Identified proteins used for database development:

In our published studies, the results showed that isopropanol depleted most of the β-conglycinin subunits, however, some glycinin isomers were observed **(Figure 1, spots 1-14).** Glycinin consists of acidic (A) and basic (B) polypeptides with five subunits G1, G2, G3, G4 and G5 **[9].** Beilinson *et al.* **[10]** identified two additional glycinin subunits in soybean variety Resnik. The five major subunits are classified into two groups based on their physical properties. Group I consists of G1 (A1aBx), G2 (A2B1a), and G3 (A1aB1b) and group II contains G4 (A5A4B3) and G5 (A3B4). In this study, isopropanol solubilized a higher proportion of group I glycinin subunits than group II subunits. In addition, we observed ten spots (spots 15-24) of soybean lectins in isopropanol extracts. It has been reported that soybean lectins (agglutinins), which are anti-nutritional and also carbohydrate-binding proteins, are present in moderate levels in soybean, peas and clover **[11].** Nine protein spots were identified as dehydrin (spots 25-33) and 14 spots (spots 34-47) were identified as maturation associated proteins. Dehydrins are a family of late embryogenesis-abundant proteins (LEA) that normally accumulate during the later stages of seed maturation and play an important role in membrane and protein stability **[12, 13]**. Various seed maturation proteins (SMPs) are also synthesized during the later stages of seed development.

Six spots (spots 48-53) were identified as the stress induced protein, SAM22 (starvation-associated message 22), which is also known as a pathogenesis related protein, PR-10. Other anti-nutritional proteins are protease inhibitors that help protect the seed from being eaten. Two of these, Kunitz trypsin inhibitor (KTI) and Bowman-Birk inhibitor (BBI), are well studied **[14].** We found eight spots of the trypsin inhibitor, subtype A (spots 54-61), nine spots of KTI (spots 62-70) and one spot of BBI (spot 71). KTI is an abundant protein that can inhibit trypsin, an important animal digestive enzyme. Bowman-Birk proteinase inhibitors (BBIs) are cys-rich protease inhibitors that have been identified in the Fabaceae, including

soybean (*G. max*) and Lima beans (*Phaseolus lunatus*) **[15].** We observed two spots (spot 72-73) of alcohol dehydrogenase (ADH) which are necessary for successful germination under low oxygen conditions **[16],** and five spots (spots 74-78) of malate dehydrogenase. We further observed four polypeptides (spots 79-83) of superoxide dismutase. Superoxide dismutases (SODs) are ubiquitous metallo-enzymes, which convert the superoxide radical to hydrogen peroxide and molecular oxygen, and thereby mitigate oxidative damage in all organisms **[17].** Additional identified protein spots included: three spots of napin type 2S albumin 1 precursor (spot 84-86); two spots of 2S albumin (spots 87-88); two spots of proteasome, subunit alpha type 5 and 6 (spots 89-90); three spots of type IIIa membrane protein (spots 91-93); four spots of class III acidic endochitinase (spots 94-97), one spot of nucleoside diphosphate kinase 1 (spot 98); one spot of triose phosphate isomerase (spot 99); one spot of dihydroorotate dehydrogenase (100); one spot of chain A soybean peroxidase (101); one spot of glyoxalase (102); one spot of dienelactone hydrolase (spot 103); and one spot of a disease resistance gene like protein (104). An additional three spots were categorized as unnamed/hypothetical proteins (105-107).

Construction of database:

In this investigation, previously published results on low abundant proteins were used to develop an online database. The Low Abundance Soybean Proteins (SoyProLow) database was created to organize protein data and to enable queries by users. SoyProLow stores relevant data relating to all 107 low abundance proteins described above and retrieves the corresponding information on that protein based on the spot number or "SpotID". SoyProLow retrieves theoretical_PI/Mr, protein identity, constituent peptides, sequence coverage, MOWSE_score, expected value, NCBI accession number and MS method.

The SoyProLow database was designed implemented and hosted using Microsoft SQL Server 2008 Enterprise Edition. Microsoft Visual Studio 2008 was used to design and launch web pages, which were programmed using ASP.NET with C# programming language. The server is running Microsoft Windows Server 2003 and Internet Information Services version 6.0 (IIS V6.0). SoyProLow stores unique "SpotID" for each protein and this enables users to query the database based on their specific requirements. Both the database and the website are on the same server at Towson University in Towson, MD, USA.

We developed a user friendly website to provide browse capabilities for all the data stored in the database. Users can search protein data through the "SpotID". All website pages were built based on a master page to create a consistent layout for all the pages in the website.

Users enter the "SpotID" and if the number is valid, their query will be successful. However, if the user inputs an incorrect "SpotID", i.e. one does not exist in the database; an error message is obtained. This is where web form validation comes into play. The purpose of web form validation is to ensure that the user has provided proper input needed to successfully complete a query. The website also validates user input by type and if the user enters a character instead of an integer number as "SpotID", an error message will occur. It is a very rich validation and the query never gets submitted if validation fails. SQL scripts were used to build a database driven search page. Integration between ASP.NET and SQL made it possible to construct powerful search capabilities. **Figure1** shows a snapshot of the actual webpage. The web-based database is publically available at the following address: http://bioinformatics.towson.edu/Soybean_low_abundance_proteins_2D_Gel_DB/Gel1.aspx.

Acknowledgement:

We thank Dr. Richard Sicher for critical review of this manuscript. Funding for this research was provided by ARS project 1245-21220-232-00D. Mention of trade name, proprietary product or vendor does not constitute a guarantee or warranty of the product by the U.S. Department of Agriculture or imply its approval to the exclusion of other products or vendors that also may be suitable.

References:

[1] Friedman M & Brandon DL, *J Agri Food Chem*. 2001 **49:** 1069 [PMID: 11312815]
[2] Liener IE, *Crit Rev Food Sci Nutri.* 1994 **34:** 31 [PMID: 8142044]
[3] Dubey H & Grover A, *Current Sci*. 2001 **80**: 262
[4] Luo J *et al. J Proteome Res*. 2009 **8**: 829 [PMID: 18778094]
[5] Wang W *et al. Electrophoresis.* 2006 **27**: 2782 [PMID: 16732618]
[6] Natarajan SS *et al. Anal Biochem.* 2005 **342**: 214 [PMID: 15953580]
[7] Xu C *et al. Phytochem*. 2006 **67**: 2431 [PMID: 17046036]
[8] Natarajan SS *et al. Anal Biochem*. 2009 **394**: 259 [PMID: 19651100]
[9] Nielsen NC *et al. Plant Cell*. 1989 **1**: 313 [PMID: 2485233]
[10] Beilinson V *et al. Theor Appl Genet.* 2002 **104**: 1132 [PMID: 12121465]
[11] Vodkin LO *et al. Plant Physiol.* 1986 **81**: 558 [PMID: 16664856]
[12] Battaglia M *et al. Plant Physiol*. 2008 **148**: 6 [PMID: 18772351]
[13] Samarah NH *et al. Crop Sci*. 2006 **46**: 2141
[14] Laskowski M & Kato I, *Annu Rev Biochem*. 1980 **49**: 593 [PMID: 6996568]
[15] Qu LJ *et al. Plant Physiol*. 2003 **133**: 560 [PMID: 12972663]
[16] Mooney BP *et al. Phytochemistry.* 2004 **65**:1733 [PMID: 15276434]
[17] Fridovich I, *J Biol Chem.* 1989 **264**: 7761 [PMID: 2542241]

Edited by P Kangueane

Citation: **Tavakolan *et al.*** Bioinformation 10(9): 599-601 (2014)