Phytochemistry 72 (2011) 1020–1027

Contents lists available at ScienceDirect

Phytochemistry

journal homepage: www.elsevier.com/locate/phytochem

A legume specific protein database (LegProt) improves the number of identified peptides, confidence scores and overall protein identification success rates for legume proteomics

Zhentian Lei, Xinbin Dai, Bonnie S. Watson, Patrick X. Zhao, Lloyd W. Sumner *

Plant Biology Division, The Samuel Roberts Noble Foundation, 2510 Sam Noble Parkway, Ardmore, OK 73401, USA

ARTICLE INFO

Article history:
Available online 23 February 2011

Keywords:
Medicago truncatula
Glycine max
Lotus japonica
Medicago sativa
Lupinus albus
Phaseolus vulgaris
Pisum sativum
Legume proteomics
Tandem mass spectrometry
Protein database

ABSTRACT

A legume specific protein database (LegProt) has been created containing sequences from seven legume species, i.e., *Glycine max*, *Lotus japonicus*, *Medicago sativa*, *Medicago truncatula*, *Lupinus albus*, *Phaseolus vulgaris*, and *Pisum sativum*. The database consists of amino acid sequences translated from predicted gene models and 6-frame translations of tentative consensus (TC) sequences assembled from expressed sequence tags (ESTs) and singleton ESTs. This database was queried using mass spectral data for protein identification and identification success rates were compared to the NCBI nr database. Specifically, Mascot MS/MS ion searches of tandem nano-LC Q-TOFMS/MS mass spectral data showed that relative to the NCBI nr protein database, the LegProt database yielded a 54% increase in the average protein score (i.e., from NCBI nr 480 to LegProt 739) and a 50% increase in the average number of matched peptides (i.e., from NCBI nr 8 to LegProt 12). The overall identification success rate also increased from 88% (NCBI nr) to 93% (LegProt). Mascot peptide mass fingerprinting (PMF) searches of the LegProt database using MALDI-TOFMS data yielded a significant increase in the identification success rate from 19% (NCBI nr) to 34% (LegProt) while the average scores and average number of matched peptides showed insignificant changes. The results demonstrate that the LegProt database significantly increases legume protein identification success rates and the confidence levels compared to the commonly used NCBI nr. These improvements are primarily due to the presence of a large number of legume specific TC sequences in the LegProt database that were not found in NCBI nr. The LegProt database is freely available for download (http://bioinfo.noble.org/manuscript-support/legumedb) and will serve as a valuable resource for legume proteomics.

1. Introduction

Legumes (Fabaceae) are one of the most economically important crop families in the world, only second to Poaceae (grass and cereals). They are planted on about 15% of the world's arable land (270–300 million hectares) and provide 33% of dietary protein and 35% of vegetable oil for the world (Graham and Vance, 2003). In the United States alone, over 77 million acres of soybeans (*Glycine max*) were cultivated in 2009 (http://quickstats.nass.usda.gov/results/959DEA59-E184-30C4-817C-0BA16752E450?pivot=short_desc) and had an estimated value of more than $31 billion (http://quickstats.nass.usda.gov/results/B5865F8D-CCC3-3E39-B363-28A7598A314C?pivot=short_desc). Legumes are important food and forages because of their ability to convert or 'fix' atmospheric nitrogen through a symbiotic interaction with Rhizobia. During this interaction, atmospheric nitrogen is reduced to ammonia which is readily incorporated in amino acid biosynthesis and ultimately results in high protein content. It was estimated that $7–10 billion worth of nitrogen is fixed by legumes annually and a proportion returned to the soil for subsequent crops. Thus, even modest use of alfalfa (*Medicago sativa*) in rotation with corn could save farmers $200–300 million in nitrogen fertilizer costs annually (Peterson and Russelle, 1991). Legumes are also a unique source of natural products such as flavonoids, isoflavonoids, alkaloids, and saponins, many of which have documented antimicrobial, pharmaceutical, and/or neutraceutical properties.

Medicago truncatula and *Lotus japonicus* have been selected as model legumes primarily due to their small genome size (each about 500 MB as compared to 1100 MB for soybean), self-fertilization, genetic transformability, and prolific nature (Barker et al., 1990; Bell et al., 2001; Cook, 1999; Trieu et al., 2000). Large numbers of tentative consensus (TC) sequences assembled from overlapping EST sequences have been generated for *M. truncatula*, *L. japonicus* and soybean (http://compbio.dfci.harvard.edu/tgi/plant.html). Further, a physical map and genome sequence of the gene-rich space of *M.*

* Corresponding author. Tel.: +1 580 224 6710; fax: +1 580 224 6692.
E-mail address: lwsumner@noble.org (L.W. Sumner).

0031-9422/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.phytochem.2011.01.026

truncatula are now available (Choi et al., 2004; Thoquet et al., 2002; Young et al., 2005). The genome sequencing of *L. japonicus* has also been conducted in Japan (Cannon et al., 2006; Sato et al., 2008). The wealth of available nucleotide sequences constitutes an important genetic resource for molecular biology as well as proteomic research in legumes.

The rapid developments in proteomics have made it a valuable tool in biology. Two-dimensional gel electrophoresis (2-DE) is a well established tool for separating proteins prior to mass spectrometry (MS) protein identifications following in-gel trypsin digestions (Gorg et al., 2000). 2-DE separates proteins based on their isoelectric point and molecular weight. Modern 2-DE is capable of resolving several thousand proteins making it still one of the highest resolution and common methods of choice in proteomics (Görg et al., 2004; Lei et al., 2005).

Proteomic approaches have been successfully employed to investigate many different legume species including soybean (for review, see (Komatsu and Ahsan, 2009)), *M. truncatula* (Bestel-Corre et al., 2002; Colditz and Braun, 2010; Imin et al., 2004; Lei et al., 2005; Mathesius et al., 2001; Soares et al., 2007; Watson et al., 2003; Zhang et al., 2006), alfalfa (Incamps et al., 2005; Watson et al., 2004), lupin (*Lupinus albus*) (Brambilla et al., 2009; Tian et al., 2009), *L. japonicus* (Dam et al., 2009), chickpea (*Cicer arietinum*) (Bhushan et al., 2007; Pandey et al., 2008), pea (*Pisum sativum*) (Bourgeois et al., 2009; Curto et al., 2006; Saalbach et al., 2002), common bean (*Phaseolus vulgaris*) (Lee et al., 2009), and white clover (*Trifolium repens*) (Wilson et al., 2002). Most of the previous legume proteomics studies have used the National Center for Biotechnology Information non-redundant (NCBI nr) protein database for protein queries and identifications. The NCBI nr protein database is a very comprehensive and large protein database, consisting of amino acid sequences translated from coding genome sequences from a vast number of different organisms and species including human, animals, plants and microorganisms.

For probability-based protein identification algorithms such as Mascot, the reported significance threshold score is proportional to the number of sequences being searched within the database because the probability of obtaining a random match increases with the size of the database. Thus, protein identification success rates are inversely proportional to the size of the database when all else is the same (Bienvenut et al., 2002). In addition, a large number of experimentally determined tentative consensus sequences and expressed sequence tags (i.e., TCs and ESTs) have not been included into the NCBI nr protein database to date. Thus, we hypothesized that a legume specific protein database that contained a comprehensive compilation of legume sequences would be highly beneficial in increasing protein identification confidence due to its smaller size and more importantly its greater legume specific and comprehensive TC, EST, and genomic content. A legume specific protein database was assembled to quantify the utility of such a database on legume protein identification confidence and success relative to the NCBI nr database. The results show that substantial increases in protein identification and confidence were achieved using LegProt relative to NCBI nr for both Mascot tandem MS/MS ion and peptide mass fingerprint (PMF) searches.

2. Results and discussion

2.1. The LegProt database

The LegProt database was created with the intent to enhance protein identification success rates and confidence levels. For this purpose, such a database should include traditional genomic and published protein sequences as well as the wealth of more recent singleton EST and TC sequences. Towards this goal, protein sequences from different legume species and different origins (i.e., genome sequences and experimentally determined sequences) were all included into the LegProt database. The LegProt database used in the present comparison was assembled from 175,787 original amino acid sequences from seven legume species (*G. max*, *L. japonicus*, *M. sativa*, *M. truncatula*, *P. sativum*, *P. vulgaris* and *L. album*) and 121 *Arabidopsis* mitochondrial proteins. These amino acid sequences, based on their origins, were classified into predicted genes on genomic sequences (37,851 sequences or 21%), published legume protein sequences (6439 sequences or 4%), and a large number of TCs and singleton ESTs (131,497 sequences or 75%). The predicted gene models were obtained from three model legumes whose genomes were partially sequenced at the time of the initial LegProt database compilation. These model legumes were *M. truncatula* (27,899 sequences), *G. max* (9424 sequences) and *L. japonicus* (9952 sequences), respectively. Published protein sequences (6439 sequences) including chloroplast and mitochondrial proteins were from several species: *L. japonicus* chloroplast proteins (247 sequences), *G. max* chloroplast proteins (317 sequences), *Arabidopsis* mitochondria proteins (121 sequences), *M. sativa* proteins (1908 sequences including 36 chloroplast proteins), *P. sativum* proteins (1942 sequences), *P. vulgaris* proteins (1223 sequences), and *L. albus* (681 sequences). These proteins were downloaded from NCBI and incorporated into the LegProt database because chloroplast and mitochondrial protein sequences are typically under-represented in ESTs and TCs due to the lack of poly(A) tails on their mRNA sequences (Slomovic et al., 2006).

The LegProt database has two major advantages compared to the NCBI nr database. First, it is legume specific. Over 99.9% of the sequences in the database were from legumes (i.e., except for the 121 *Arabidopsis* mitochondrial proteins). These legume specific sequences are important in legume proteomics as legumes differ from other plants in their ability to form symbiosis with Rhizobia. This unique nodulation process requires some legume-specific proteins not found in other organisms. For example, it has been shown that a large number of nodule-specific cysteine rich proteins and glycine rich proteins are only present in legumes, but not in other plants (Alunni et al., 2007; Mergaert et al., 2003). Blocking the secretory pathway of these proteins results in abnormal bacteroid and symbiosome development, hence impairment of nitrogen fixation (Wang et al., 2010). In addition, species-specific preferences for amino acid residues such as Leu, Cys, Asp, Thr, Ser, and to lesser extent Glu, Gln, His and Met have been observed for taxonomically diverse organisms (Dumontier et al., 2002). These facts indicate that a legume specific protein database would be beneficial in legume proteomics. Because of its legume specific nature, the database size is significantly smaller than NCBI nr. The LegProt database used in this work had 175,787 original sequences while NCBI nr (version of September 24, 2010) contained a total of 11,894,394 sequences. Although searching against NCBI nr can be restricted to green plants (Viridiplantae, 839,501 sequences), *Arabidopsis* (63,177 sequences), *Oryza sativa* (rice, 134,475 sequences), or other green plants (642,608 sequences), searching specifically against legume species is not available. The larger size of the NCBI nr database leads to higher significance threshold scores for probability based searches, and therefore may lead to lower identification success rates.

The second major advantage of the LegProt database is that it contains a large number of experimentally determined sequences (i.e., TCs and singleton ESTs), accounting for 76% of the sequences in the database. These sequences are assembled directly from mRNA sequences and therefore represent the corresponding protein sequences more accurately than those predicted from genome sequences. Unfortunately, these sequences have not been included in the NCBI nr protein database to date. Instead, NCBI nr contains predominantly translations of all GenBank DNA coding sequences. Thus, identification of these legume proteins through queries of the