Computational Molecular Biology

2014, Vol.4, No.1, 1-17 http://cmb.sophiapublisher.com

BioPublisher©

Research Report Open Access

PlantSecKB: the Plant Secretome and Subcellular Proteome KnowledgeBase

Gengkon Lum[1,3], John Meinken[1], Jessica Orr[2], Stephanie Frazier[2], Xiang Jia Min[2,3]

1. Department of Computer Science and Information Systems, Youngstown State University, OH 44555, USA

2. Department of Biological Sciences, Youngstown State University, Youngstown, OH 44555, USA

3. Center for Applied Chemical Biology, Youngstown State University, Youngstown, OH 44555, USA

Corresponding Author email: xmin@ysu.edu; Author

Computational Molecular Biology, 2014, Vol.4, No.1 doi: 10.5376/cmb.2014.04.0001

Abstract Prediction and curation of protein subcellular locations is essential for protein functional annotation. We developed the Plant Secretome and Subcellular Proteome KnowledgeBase (PlantSecKB) for the plant research community to access and curate plant protein subcellular locations, with a focus on secreted proteins. The database is constructed with all the available plant protein data retrieved from the UniProtKB database and plant protein sequences predicted from EST data assembled by the PlantGDB project. The database contains information collected from three sources: (1) subcellular locations that were curated or computationally predicted in the UniProtKB; (2) subcellular locations and features predicted by eight computational tools; (3) secreted proteins that were curated from recent literature. The categories of subcellular locations include secretome, mitochondria, chloroplast, cytosol, cytoskeleton, endoplasmic reticulum, Golgi apparatus, lysosome, peroxisome, nucleus, vacuole, and plasma membrane. The data can be searched by using UniProt accession number or ID, GenBank GI or RefSeq accession number, gene name, and keywords. Species specific secretome and subcellular proteomes can be searched and downloaded into a FASTA file. BLAST is available to allow users to search the database based on protein sequences. Community curation for subcellular locations of plant proteins is also supported. A primary analysis revealed that monocots and dicots had a similar proportion of secretomes, and monocots had a significantly higher proportion of proteins distributed to mitochondria (both membrane and non-membrane) and chloroplast membrane, while dicots had significantly more proteins distributed to cytosol and nucleus. This database aims to facilitate plant protein research and is available at http://proteomics.ysu.edu/secretomes/plant.php.

Keywords Computational prediction; Expressed sequence tags; Plant secreted protein; Secretome; Signal peptide; Subcellular location; Subcellular proteome

Introduction

Plants are the main contributors to the production of biomass including carbohydrates, proteins, lipids, cellulose and other biomaterials. Plant proteins including enzymes, regulatory and structural proteins play important biological roles in regulating plant growth and development. Plant proteins are synthesized within a cell and transported to different subcellular locations including extracellular space or matrix to perform their biological functions. This process often is called protein sorting or targeting (Foresti and Denecke, 2008; Rose and Lee, 2010).

Plant cells contain a cell wall, a plasma membrane, choloroplasts, mitochondria, a large vacuole, a nucleus, endoplasmic reticulum (ER), a Golgi apparatus, peroxisomes, cytosol, etc. Membrane proteins can be embedded or attached to plasma membrane, organelle membrane or endomembrane systems.

Identification and analysis of protein subcellular locations in eukaryotes is one of the important subjects for annotating a proteome. In a plant species, proteins secreted to the extracellular space or matrix, which includes the cell wall, are collectively called a "secretome" (Agrawal et al., 2010; Lum and Min, 2011a).

Preferred citation for this article:

Min et al., 2014, PlantSecKB: the Plant Secretome and Subcellular Proteome KnowledgeBase, Computational Molecular Biology, Vol.4, No.1 1-17 (doi: 10.5376/cmb.2014.04.0001)

Received: 04 Dec., 2013 | Accepted: 24 Dec., 2013 | Published: 24 Feb., 2014

The term secretome was first introduced by Tjalsma et al. (2000) to denote the complete set of proteins in *Bacillus subtilis* processed by the secretory pathway, which included protein secreted to the extracellular space and also proteins involved in the pathway. However, recently it was more often limited, as in this work, to represent only the secreted, extracellular portion - including cell wall proteins - of the proteome (e.g., Greenbaum et al., 2001; Hathout, 2007; Bouws et al., 2008; Agrawal et al., 2010; Lum and Min, 2011b). A plant secretome consists of primarily cell wall proteins, proteins involved in cell wall metabolism, and extracellular enzymes and signal molecules involved in defense of pathogens (Isaacson and Rose, 2006; Kamoun, 2009; Lum and Min, 2011a). Secreted enzymes, particularly hydrolases such as α-amylase and α-glucosidases, have been well studied using germinating barley seeds as a model system. These hydrolases were synthesized in the aleurone layer and secreted into the endosperm to break down starch and other storage reserves (Ranki and Sopanen, 1984; Jones and Robinson, 1989; Finnie et al., 2011 for review). Recently, advances in proteomic analytic techniques along with the complete sequencing of *Arabidopsis thaliana* and *Oryza sativa* genomes resulted in many secreted proteins, including the cell wall proteome, being identified (Boudart et al., 2007; Agrawal et al., 2010; Lum and Min, 2011a). These identified secreted proteins mainly consist of cell wall proteins in Arabidopsis (see Jamet et al., 2008 for review) and some enzymes such as GLP1 involved in pathogen defense (Oh et al., 2005). Using a leaf or seed cell suspension culture, secreted proteins were identified with 2D-gel electrophoresis coupled with liquid chromatography mass spectrometry analysis in rice, Medicago and sorghum (Jung et al., 2008; Kusumawati et al., 2008; Cho et al., 2009; Ngara and Ndimba, 2011). A large number of secreted proteins were also identified from root exudates using aseptically grown seedlings of rice and Arabidopsis (Shinano et al., 2011; De-la-Pena et al., 2010). Experimental systems, analytical techniques, and related bioinformatics tools used for plant secretome study were recently comprehensively reviewed (Agrawal et al., 2010; Meinken and Min, 2012; Alexandersson et al., 2013; Kraus et al., 2013; Caccia et al., 2013).

Classical eukaryotic secreted proteins contain a secretory signal peptide at the N-terminus that directs proteins to the rough ER for completing protein synthesis and then transports them to the Golgi complex for protein targeting (von Heijne, 1990). The signal peptide, typically 15~30 amino acids long, is often cleaved off during translocation across the endomembrane systems. Classical secreted proteins can be computationally predicted relatively accurately (Min, 2010). Recently we analyzed all manually curated and annotated secreted plant proteins in the UniProtKB/Swiss-Prot dataset and found 87% of them could be predicted to have a signal peptide by all three predictors used (Lum and Min, 2011a). The accuracy of secretome prediction could be further improved by using a new version of SignalP (SignalP 4.0) combined with other tools including TMHMM for identifying transmembrane proteins and PS-Scan for identifying ER luminal proteins (Min, 2010; Melhem et al., 2013).

With improvements in sequencing technology and the reduced cost of sequencing, the genomes of more and more plant species are being completely sequenced. Currently there are 32 land plants with complete or draft genome sequences available and 73 land plant species with genome sequencing in progress (http://www.ncbi.nlm.nih.gov/genomes/static/gpstat.html). There are also assembled expressed sequence tag (EST) data in plants available for identifying potential genes encoding secreted proteins in more than 200 species (PlantGDB, http://www.plantgdb.org/prj/ESTCluster/) (Duvick et al., 2008). As a result of genome sequencing, the number of protein sequences available is increasing rapidly.

In addition to the classical secreted proteins, a large number of leadless, non-classical, secreted proteins (LSP), i.e. not having a secretory signal peptide, have been identified in plants (Jung et al., 2008; Agrawal et al., 2010; Ding et al., 2012 for review). These proteins have not been curated in the UniProtKB. Therefore there is a need to have a central knowledgebase providing plant protein subcellular locations for the

plant research community to access the available information and deposit experimental evidence for newly characterized proteins. In order to provide such a central plant secretome related resource portal, we developed the Plant Secretome and Subcellular Proteome KnowledgeBase (PlantSecKB) (http://proteomics.ysu.edu/secretomes/plant.html), which includes predicted and manually curated protein subcellular locations from plant proteomes as well as predicted proteins from EST data in plants. Though our focus is on plant secretomes, the information on proteins located in other subcellular locations is also provided. A tool for supporting community manual curation of plant protein subcellular locations can be accessed through the database interface.

1 Methods of Database Construction

1.1 Data collection

PlantSecKB was constructed primarily with the sequence data obtained from two sources: plant protein sequences extracted from UnitProtKB (2013-04 Release) (http://www.uniprot.org/) and protein sequences predicted from assembled EST data compiled by the PlantGDB project (http://www.plantgdb.org/prj/ESTCluster/). The proteins predicted from the recently sequenced sacred lotus (*Nelumbo nucifera Gaertn*.) genome were also integrated into this database (Ming et al., 2013; Lum et al., 2013). Protein sequences in the EST data were predicted using the OrfPredictor tool (http://proteomics.ysu.edu/tools/OrfPredictor.html) with BLASTX input against the UniProt/Swiss-Prot database, and TargetIdentifier (http://proteomics.ysu.edu/tools/TargetIdentifier.html) was used to examine if an EST was full-length (Min et al., 2005a, 2005b).

1.2 Computational methods for prediction of protein subcellular locations

The software tools used in this study include SignalP 3.0 and 4.0, TargetP, Phobius, WoLF PSORT, TMHMM, PS-Scan, and FragAnchor. The website links for these tools and related references can be found in our website (http://proteomics.ysu.edu/tools/subcell.html). Except FragAnchor, we used the standalone tools installed on a local Linux system for data processing. The commands for how to run them often can be found in the "readme" page in each downloaded package and were summarized by Lum and Min (2013). In brief, SignalP 4.0 was used for secretory signal peptide prediction (Petersen et al., 2011). However, we also included prediction information from SignalP 3.0 (Bendtsen et al., 2004b) as it provides more accurate cleavage site prediction than SignalP 4.0 (Petersen et al., 2011). Phobius is a combined signal peptide and a transmembrane topology predictor (Käll et al., 2007). TargetP predicts the presence of any signal sequences such as signal peptide (SP), chloroplast transit peptide (cTP) or mitochondrial targeting peptide (mTP) in the N-terminus (Emanuelsson et al., 2000; Emanuelsson et al., 2007). TMHMM uses a hidden Markov model (HMM) to predict the presence and topology of transmembrane helices and their orientation to the membrane (in/out) (Krogh et al., 2001). PS-Scan was used to scan the PROSITE database (http://www.expasy.org/tools/scanprosite/) for removing ER targeting proteins (Prosite: PS00014) (de Castro et al., 2006; Sigrist et al., 2010). FragAnchor was used to identify the glycosylphosphatidyinositol (GPI) anchored proteins (GAP) from the proteins which were predicted as containing a signal peptide by SignalP 4.0 (Poisson et al., 2007). WoLF PSORT predicts multiple subcellular locations including choloroplast, cytosol, cytoskeleton, ER, extracellular (secreted), Golgi apparatus, lysosome, mitochondria, nuclear, peroxisome, plasma membrane, and vacuolar membrane (Horton et al., 2007). The default parameters for eukaryotes or plants, if available, were used for all the programs. Our previous evaluation found that including WoLF PSORT for plant secretome prediction resulted in an accuracy decrease due to a significant decrease in the prediction sensitivity (Min, 2010). Thus, it was not used for secretome prediction but only for prediction of some other subcellular locations.

For the assignment of a subcellular location of a protein, the UniProtKB annotated subcellular location and our manual curation take precedence over computational prediction. Thus, only proteins not having an annotated subcellular location are subjected to computational assignment of their subcellular locations. The information produced by all the tools,

however, is available for all plant proteins. Some of the proteins may have more than one subcellular location. The following criteria are applied for computational classification of protein subcellular locations:

Membrane proteins: A protein predicted to contain one or more transmembrane domains by TMHMM is classified as a membrane protein. However, if there is only one transmembrane domain predicted and that is located within the N-terminus 70 amino acids, and also a signal peptide is predicted by SignalP 4.0, this protein is not counted as a membrane protein.

Chloroplast proteins: A protein predicted as "C" (for chloroplast) for subcellular location by TargetP is classified as a chloroplast protein. If it is also classified as a membrane protein, then it is further classified as chloroplast membrane protein.

Mitochondrial proteins: A protein predicted as "M" (for mitochondrial) for subcellular location by TargetP is classified as a mitochondrial protein. If it is also classified as a membrane protein, then it is further classified as mitochondrial membrane protein.

ER proteins: Proteins predicted to contain a signal peptide by SignalP 4.0 and an ER target signal (Prosite: PS00014) by PS-Scan were treated as luminal ER proteins.

Complete secretomes: A secretome is all secreted proteins from a species. Only proteins that are predicted to have a secretory signal peptide by all three predictors - SignalP 4.0, Phobius, and TargetP - and that are not classified as any of the above categories are included in the secreteome. However, proteins that are not classified as any of the above categories and are predicted to have a signal peptide by one or two of the predictors are assigned as "weakly likely secreted" or "likely secreted" as our previous evaluation revealed that a signal peptide in some annotated secreted proteins can only be detected by one or two predictors (Lum and Min, 2011a). Using all three predictors, which increases the specificity of secretome prediction, improves prediction accuracy (Min, 2010; Melhem et al., 2013). All manually curated secreted and extracellular proteins are included in the complete secretomes.

Curated secreted proteins: This category includes proteins which are annotated to be "secreted" or "extracellular" or "cell wall" in the subcellular location from the UniProtKB/Swiss-Prot data set which are "reviewed". It also includes manually collected secreted proteins from recent literature by our curators.

GPI-anchored proteins: Signal peptide containing proteins that were predicted to have a GPI anchor by FragAnchor were further classified as GPI-anchored proteins. Protein sequences predicted to have a signal peptide and a GPI anchor may attach to the outer leaflet of the plasma membrane or be secreted becoming components of the cell wall. These proteins are involved in signaling, adhesion, stress response, and cell wall remodeling or play other roles in growth and development (Borner et al., 2002; Borner et al., 2003; Gillmor et al., 2005; Simpson et al., 2009).

Proteins in other subcellular locations: Other subcellular locations including cytosol (cytoplasm), cytoskeleton, Golgi apparatus, lysosome, nucleus, peroxisome, plasma membrane and vacuole were predicted by WoLF PSORT.

1.3 Computational prediction accuracies of protein subcellular locations

The prediction methods we used above were developed based on our previous evaluation of computational tools (Min, 2010; Meinken and Min, 2012; Melhem et al., 2013). To estimate the prediction accuracies of our methods for each subcellular location we used two datasets (Table 1). Dataset A consists of 15 028 proteins. This dataset contains proteins from the UniProtKB/Swiss-Prot dataset with a curated subcellular location. Proteins having multiple subcellular locations or labeled as "fragment" were excluded. Dataset B consist of 6 908 proteins which were generated from Dataset A after excluding entries having a term of "by similarity" or "probable" or "predicted" in subcellular location annotation. In comparing with other methods using a single tool, our method - i.e. using a combination of multiple tools

including SignalP 4.0, TargetP, and Phobius for secretory signal peptide prediction and PS-Scan for removing ER proteins and TMHMM for removing membrane proteins - significantly improved the prediction accuracy for secretomes (Min, 2010; Meinken and Min, 2012). For secretome prediction our method had reached a sensitivity of 91.1%, a specificity of 98.7%, and a Mathews' correlation coefficient (MCC) of 88.5% for dataset A; and a sensitivity of 76.8%, a specificity of 98.9%, and a MCC of 74.5% for dataset B, which were much better than using WoLF PSORT or MultiLoc alone (Meinken and Min, 2012). Thus the prediction of secreted proteins is relatively reliable. The accuracies for predicting other subcellular locations still need to be improved.

Table 1 Evaluation of prediction accuracies of plant protein subcellular locations

Subcellular location	Dataset A (total 15028)					Dataset B (total 6908)				
	Total positives	Total negatives	Sn (%)	Sp (%)	MCC (%)	Total positives	Total negatives	Sn (%)	Sp (%)	MCC (%)
Secreted	1485	13543	91.1	98.7	88.5	263	6645	76.8	98.9	74.5
Mitochondrial	919	14109	65.2	82.6	28.4	402	6506	61.4	77.5	21.1
Chloroplast	8124	6904	27.5	90.9	23.5	4918	1990	28.2	90.7	20.4
ER	256	14772	22.3	100.0	46.0	87	6821	18.4	100.0	42.7
Cytosol	77	14951	61.0	78.9	7.0	23	6885	52.2	75.3	3.7
Golgi Apparatus	260	14768	1.5	99.9	6.3	54	6854	0.0	100.0	-0.2
Peroxisome	136	14892	24.3	99.7	31.6	52	6856	13.5	99.5	15.0
Nucleus	3099	11929	62.2	89.2	50.7	788	6120	68.8	85.5	42.7
Plasma Membrane	91	14937	35.2	95.1	10.7	14	6894	21.4	98.9	8.5
Vacuole	273	14755	5.1	99.0	5.5	121	6787	2.5	99.8	6.8
Cytoskeleton	305	14723	13.8	99.7	24.3	186	6722	21.0	99.7	36.0

Note: Sn: sensitivity; Sp: specificity; MCC: Mathews' correlation coefficient

1.4 Manual curation and community annotation

PlantSecKB supports community curation of subcellular locations of plant proteins based on published experimental evidence. A submission tool was developed for the community to provide subcellular location annotation of a protein and a literature source to support its annotation. After our curator's validation, these data are also incorporated into the database. Currently, based on published experimental evidence, we have manually curated 736 total secreted proteins from rice (Jung et al., 2008; Cho et al., 2009; Cho and Kim, 2009; Chen et la., 2009; Zhang et al., 2009; Shinano et al., 2011), Arabidopsis (De-la-Pena et al., 2010), and sorghum (Ngara et al., 2011). Manual curation is an ongoing process, thus more secreted proteins will be manually curated and integrated into the database in the future from the community and our curators. The information from computational prediction, UniProtKB annotation and manual curation is integrated and displayed on the annotation page (Figure 1). The annotated entries are linked to the tools used, UniProtKB, the RefSeq database and PubMed in the National Center for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov/).

2 Overview of the Database Content and Tools

2.1 Data and tool access

The PlantSecKB is accessed through the database web interface at http://proteomics.ysu.edu/secretomes/plant.php. The interface provides various utilities for searching proteins obtained from UnitProtKB, links to BLAST, an EST data search page, and the community annotation page (Figure 1). All plant proteins obtained from UniProt can be searched using UniProt accession number (AC) or ID, gene name, key word(s) in protein function or species. Sub-proteomes including curated secreted proteins, complete secretome,

mitochondrial membrane proteins, ER proteins, and others can be searched or downloaded by selecting species from a species list for those having greater than 1 000 protein sequences. Species having fewer than 1 000 protein entries can be searched by inputting a species name. The BLAST utility can be accessed through a link on the interface for searching all plant proteins or secretomes. The interface also provides a link to an EST data search page. EST data can be searched using EST identifier, keyword(s), species or BLAST.

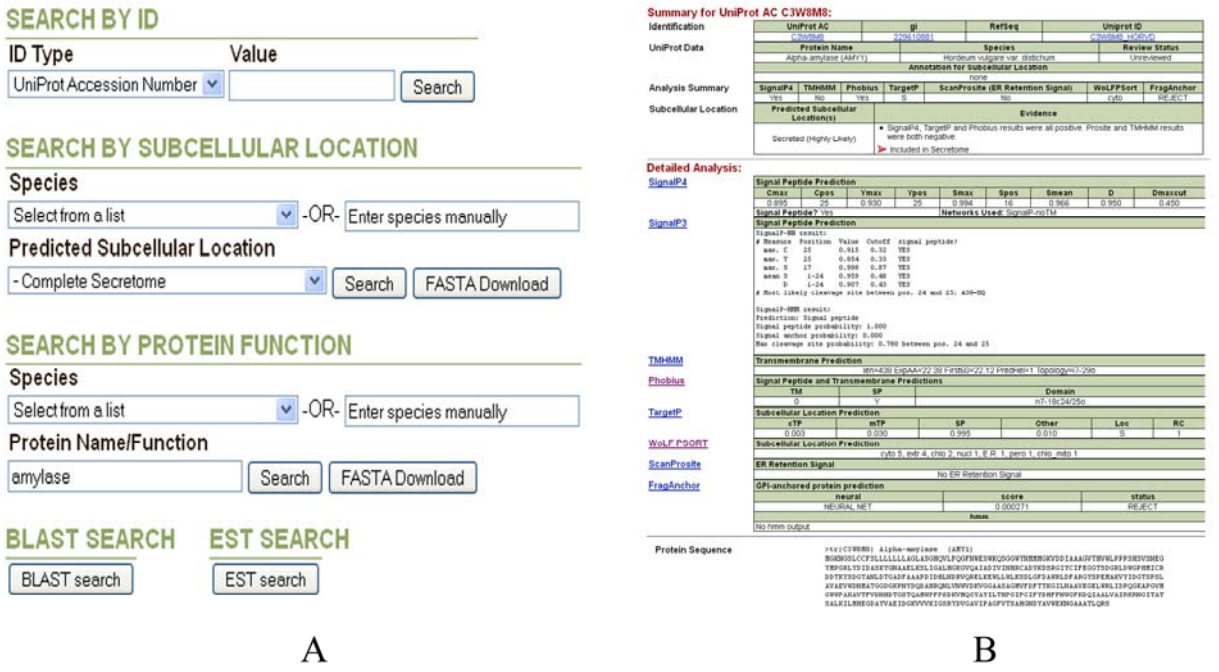

Figure 1 Overview of the PlantSecKB user interface and annotation page. (A) User interface. UniProt accession number, keywords or species can be used to search the database. Secretomes or other subcellular proteomes can be searched or downloaded. The user interface provides links to BLAST utility, EST database, and the curation submission form. (B) A page to display information of subcellular annotation, prediction, and sequence of a protein. A barley alpha-amylase is used as an example

The annotation display page for each UniProt protein contains information obtained from the following three sources: (1) the features predicted using computational approaches using the seven programs mentioned above; (2) subcellular locations annotated in UniProtKB; and (3) our manual curation with experimental evidence obtained from recent literature. The overview of the database features is shown in Figure 1. Manually curated secreted proteins consist of proteins retrieved from UniProtKB/Swiss-Prot with subcellular locations labeled as “reviewed”, as well as proteins curated by our curators. The curated proteins from internal curation and the community are supported with experimental evidence for their subcellular location annotation and related literature. The annotation page also contains the primary protein sequence (Figure 1).

EST data annotation contains the primary EST sequence, predicted protein peptide sequence using OrfPredictor (Min et al., 2005a), functional annotation based on BLASTX, prediction of completeness of the open reading frame using TargetIdentifier (Min et al., 2005b), and related information generated with the tools for subcellular location prediction based on predicted protein sequences. As EST data may contain errors introduced in sequencing and assembling, caution needs to be taken when using the data. Nevertheless, EST information provided in the database will be useful for data mining and designing experiments for further examining the gene function and subcellular locations of encoded proteins.

2.2 Data summary

PlantSecKB contains a total of 1 415 921 protein sequences including 33 643 entries from the

UniProt/Swiss-Prot dataset (curated and reviewed) and 1 355 593 from UniProt-TrEMBL (unreviewed) with an additional 26 685 proteins predicted from the newly sequenced genome of sacred lotus (Ming et al., 2013; Lum et al., 2013). The main categories of subcellular proteomes for species having more than 7 000 entries are summarized in Table 1. Curated secreted proteins, ER proteins and lysosome proteins are not listed in Table 1. There were only 7 lysosome proteins in *A. thaliana* identified and no lysosome proteins were predicted in other species. There are a total of 2 774 curated secreted proteins, which are mainly obtained from *A. thaliana* and *O. sativa* subsp. japonica with 1 247 and 559 entries, respectively. It should be noted that the number of total protein entries in a species is the number collected in the UniProtKB, which can be greater than a complete or reference genome, as there are some redundancies or duplicates in some protein entries. For example, *O. sativa* subspecies japonica has 99 984 entries in PlantSecKB and only 63 544 entries in its complete proteome set, and *A. thaliana* has 53 847 entries in PlantSecKB and only 31 908 entries in the complete proteome set in UniProtKB (http://www.uniprot.org/taxonomy/complete-proteomes).

An overall trend observed is that plants with relatively small proteome sizes have a relatively small number and a relatively lower proportion of secreted proteins, such as in single-celled green algae. For example, *Osterococcus* species has less than 100 secreted proteins predicted (1.2%), and moss (*Physcomitrella patens*) has 781 secreted proteins predicted (2.9%) (Table 2). On average the secretome accounts for about 4.0%~7.5% of the proteome in monocot and dicot plants based on our prediction estimations. The secretome percentages reported in this study are slightly lower than we reported previously. This is due to the fact that our previous study used SignalP 3.0, whereas this study used SignalP 4.0 which has a higher specificity (Lum et al., 2013; Petersen et al., 2011).

The average predicted proteome sizes and distributions of subcellular proteomes are summarized in Table 3 using 9 species or subspecies in each category of green algae, monocot and dicot plants listed in Table 2. *Lotus japonicus*, a dicot, was the only species not used for this analysis due to incompleteness of its proteome. The average predicted proteome size is much smaller in green algae, thus each subcellular proteome consists of a smaller number of proteins (Table 3). Comparing monocots and dicots, the distribution percentages of secreted proteins, chloroplast membrane proteins, vacuolar proteins, and plasma membrane proteins were not significantly different. However, monocots had a significantly higher proportion of proteins predicted as mitochondria (both membrane and non-membrane) and chloroplast membrane, and dicots had significantly more proteins predicted as cytosol and nucleus (Table 3). Whether these observed differences in subcellular proteome distributions between monocots and dicots are caused by computational tools or are real with biological or evolutionary significances needs further investigation.

Table 3 Comparison of subcellular proteome distribution in green algae, monocot and dicot plants

			Mitochondiral		Chloroplast				Plasma	
	Proteome	Secretome (%)	Membrane (%)	Non-membrane (%)	Membrane (%)	Non-membrane (%)	Cytosol (%)	Vacuole (%)	Membrane (%)	Nucleus (%)
Green algae	10371	284 (2.7)	286 (2.8)	1975 (19.0)	201 (1.9)	1284 (12.4)	1933 (18.8)	83 (0.8)	341 (3.3)	1567(145)
Monocot	43653	2667 (6.1)	834 (1.9)	7140 (16.4)	702 (1.6)	6304 (14.4)	6822 (15.6)	381 (0.9)	1699 (3.9)	7947(182)
Dicot	45715	2645 (5.8)	562 (1.2)	5098 (11.2)	712 (1.6)	5122 (11.2)	8600 (18.8)	459 (1.0)	2180 (4.8)	10342(226)
T-test	ns	ns	***	***	ns	***	***	ns	ns	***

Note: T-test was used to compare the subcellular proteome (%) distribution in monocots and dicots. ns: not significant; ***: highly significant ($t < 0.001$)

3 Comparative Analysis of Secretomes

Complete comparative evolutionary analyses of plant secretomes or other sub-proteomes were beyond the scope of this study. However, as complete secretome or other sub-proteome sequences can be downloaded directly from our database, it would facilitate further detailed comparative study of these sub-proteomes in different species. As an example, we performed a comparative analysis of secretomes using a set of representative plants including three monocots (*Brachypodium distachyon*, *Oryza sativa* subsp. japonica, *Zea mays*), three dicots (*Arabidopsis thaliana*, *Populus trichocarpa*, *Solanum lycopersicum*), and two mosses (*Physcomitrella patens* subsp. patens, *Selaginella moellendorffii*) (Table 4 and Table 5). We used the blastclust tool in the BLAST package with a cutoff of 95% identities in the aligned pair to remove or reduce redundancy. Thus non- or less redundant secreteomes were used for comparisons. To provide an overview of the functionalities of secretomes in plants, we carried out Gene Ontology (GO) analysis of representative secretomes of the 8 selected plant species. The secretomes were used to search the UniProt/Swiss-Prot dataset with BLASTP with a cutoff E-value of 1e-10. GO information was retrieved from UniProt ID mapping data (http://www.uniprot.org/downloads) and analyzed using GO SlimViewer with plant specific GO terms (McCarthy et al., 2006). Comparison of GO biological process and molecular function classification of secretomes of the selected species was summarized in Table 4. Plant secreted proteins are involved in more than 40 different biological processes including metabolic and catabolic processes, response to stress and biotic or abiotic stimulus, carbohydrate, lipid and protein metabolic processes, multicellular organismal development, etc. Molecular function classification revealed that plant secretomes consist of a large number of hydrolases (~30%) and tranferases (7%~9%), and that a large proportion of them have various binding activity (~40%) or catalytic activity (12%~15%). It should be noted that GO classification was only an estimate of the distributions of each category as many secreted proteins have not been classified in GO.

Table 4 Gene Ontology classification of secreted proteins in different plant species

(a) Biological Process		*At (%)*	*Pt (%)*	*Sl (%)*	*Bd (%)*	*Osj (%)*	*Zm (%)*	*Pp (%)*	*Sm (%)*
GO:0008152	metabolic process	673 (16)	379 (21)	439 (22)	393 (20)	544 (20)	429 (20)	155 (23)	282 (21)
GO:0006950	response to stress	579 (14)	170 (9)	200 (10)	188 (10)	260 (9)	188 (9)	59 (9)	99 (7)
GO:0009056	catabolic process	386 (9)	182 (10)	242 (12)	200 (10)	269 (10)	216 (10)	71 (10)	137 (10)
GO:0009607	response to biotic stimulus	353 (9)	61 (3)	65 (3)	49 (3)	65 (2)	54 (3)	16 (2)	29 (2)
GO:0005975	carbohydrate metabolic process	313 (8)	156 (9)	190 (9)	183 (9)	247 (9)	165 (8)	56 (8)	97 (7)
GO:0007275	multicellular organismal development	161 (4)	64 (4)	74 (4)	78 (4)	120 (4)	93 (4)	30 (4)	69 (5)
GO:0016043	cellular component organization	150 (4)	66 (4)	65 (3)	71 (4)	121 (4)	75 (4)	19 (3)	46 (3)
GO:0019538	protein metabolic process	143 (3)	98 (5)	90 (4)	98 (5)	109 (4)	91 (4)	40 (6)	71 (5)
GO:0006629	lipid metabolic process	140 (3)	65 (4)	68 (3)	72 (4)	102 (4)	83 (4)	16 (2)	61 (4)
GO:0009628	response to abiotic stimulus	111 (3)	39 (2)	39 (2)	56 (3)	82 (3)	71 (3)	14 (2)	36 (3)
GO:0007165	signal transduction	107 (3)	29 (2)	33 (2)	28 (1)	44 (2)	30 (1)	8 (1)	16 (1)
GO:0000003	reproduction	99 (2)	52 (3)	52 (3)	68 (4)	102 (4)	68 (3)	17 (2)	44 (3)
GO:0006810	transport	89 (2)	56 (3)	48 (2)	36 (2)	65 (2)	43 (2)	10 (1)	32 (2)
GO:0009058	biosynthetic process	86 (2)	66 (4)	70 (3)	62 (3)	102 (4)	89 (4)	40 (6)	60 (4)
GO:0030154	cell differentiation	86 (2)	16 (1)	20 (1)	23 (1)	44 (2)	23 (1)	8 (1)	17 (1)
	others	636 (15)	316 (17)	309 (15)	322 (17)	505 (18)	385 (18)	125 (18)	268 (20)
	total	4112	1815	2004	1927	2780	2103	684	1364

(b) Molecular Function		At (%)	Pt (%)	Sl (%)	Bd (%)	Osj (%)	Zm (%)	Pp (%)	Sm (%)
GO:0016787	hydrolase activity	649 (32)	328 (23)	380 (29)	398 (29)	533 (24)	362 (28)	114 (28)	243 (29)
GO:0005488	binding	595 (29)	435 (31)	408 (31)	434 (32)	711 (33)	407 (31)	139 (34)	263 (31)
GO:0003824	catalytic activity	249 (12)	186 (13)	158 (12)	194 (14)	272 (12)	169 (13)	59 (15)	115 (14)

Continuing Table 4

(b) Molecular Function		At (%)	Pt (%)	Sl (%)	Bd (%)	Osj (%)	Zm (%)	Pp (%)	Sm (%)
GO:0016740	transferase activity	135 (7)	122 (9)	107 (8)	97 (7)	191 (9)	116 (9)	27 (7)	75 (9)
GO:0000166	nucleotide binding	92 (4)	103 (7)	82 (6)	74 (5)	166 (8)	85 (6)	21 (5)	51 (6)
GO:0030234	enzyme regulator activity	61 (3)	29 (2)	53 (4)	28 (2)	42 (2)	23 (2)	2 (0)	2 (0)
GO:0005102	receptor binding	57 (3)	11 (1)	7 (1)	10 (1)	17 (1)	10 (1)	2 (0)	8 (1)
GO:0016301	kinase activity	51 (2)	72 (5)	55 (4)	45 (3)	106 (5)	49 (4)	13 (3)	40 (5)
GO:0004871	signal transducer activity	43 (2)	14 (1)	11 (1)	13 (1)	23 (1)	13 (1)	3 (1)	9 (1)
GO:0030246	carbohydrate binding	41 (2)	33 (2)	24 (2)	37 (3)	54 (2)	29 (2)	8 (2)	14 (2)
GO:0008289	lipid binding	27 (1)	19 (1)	26 (2)	18 (1)	23 (1)	14 (1)	1 (0)	7 (1)
	others	47 (2)	44 (3)	22 (2)	16 (1)	43 (2)	32 (2)	15 (4)	14 (2)
	total	2047	1396	1333	1364	2181	1309	404	841

Note: At: *Arabidopsis thaliana*; Pt: *Populus trichocarpa*; Sl: *Solanum lycopersicum*; Monocots - Bd: *Brachypodium distachyon*; Osj: *Oryza sativa* (subsp. japonica); Zm: *Zea mays*. Mosses - *Physcomitrella patens* (subsp. patens); Sm: *Selaginella moellendorffii*

Table 5 Comparison of protein families in secretomes of representative plant species

	Dicots			Monocots			Mosses			
Pfam ID	*At*	*Pt*	*Sl*	*Bd*	*Osj*	*Zm*	*Pp*	*Sm*	Pfam name	Pfam discription
pfam00657	90	61	52	51	76	48	9	42	Lipase_GDSL	GDSL-like Lipase/Acylhydrolase
pfam00141	73	58	82	119	151	91	22	51	peroxidase	Peroxidase
pfam04043	63	38	27	26	40	28	0	0	PMEI	Plant invertase/pectin methylesterase inhibitor
pfam05617	56	10	5	3	5	2	0	0	Prolamin_like	Prolamin-like
pfam00450	54	30	42	41	57	23	7	13	Peptidase_S10	Serine carboxypeptidase
pfam01095	52	24	37	16	26	13	5	5	Pectinesterase	Pectinesterase
pfam05938	52	13	13	0	0	0	2	0	Self-incomp_S1	Plant self-incompatibility protein S1
pfam00295	49	30	35	29	30	32	1	4	Glyco_hydro_28	Glycosyl hydrolases family 28
pfam01657	45	22	5	9	31	5	0	8	Stress-antifung	Salt stress response/antifungal
pfam00026	41	30	38	36	65	41	3	6	Asp	Eukaryotic aspartyl protease
pfam00332	37	25	22	24	43	28	3	11	Glyco_hydro_17	Glycosyl hydrolases family 17
pfam00190	35	40	38	35	56	25	12	25	Cupin_1	Cupin
pfam00722	34	22	28	26	34	29	8	10	Glyco_hydro_16	Glycosyl hydrolases family 16
pfam00232	33	11	12	21	42	9	2	12	Glyco_hydro_1	Glycosyl hydrolase family 1
pfam00234	31	16	37	20	42	34	1	2	Tryp_alpha_amyl	Protease inhibitor/seed storage/LTP
pfam01357	30	19	26	54	70	49	14	9	Pollen_allerg_1	Pollen allergen
pfam00082	30	18	44	40	39	22	1	18	Peptidase_S8	Subtilase family
pfam03080	29	2	16	5	25	9	0	6	DUF239	Domain of unknown function (DUF239)
pfam01190	27	24	15	28	47	24	1	11	Pollen_Ole_e_I	Pollen proteins Ole e I like
pfam00112	27	19	23	26	47	27	10	11	Peptidase_C1	Papain family cysteine protease
pfam05498	27	10	7	10	12	14	0	1	RALF	Rapid ALkalinization Factor (RALF)
pfam01565	26	32	21	17	15	8	1	22	FAD_binding_4	FAD binding domain
pfam07983	26	7	7	9	14	17	0	2	X8	X8 domain
pfam07732	24	40	28	32	43	20	5	8	Cu-oxidase_3	Multicopper oxidase
pfam00149	24	8	12	14	17	10	6	8	Metallophos	Calcineurin-like phosphoesterase
pfam07333	24	0	0	0	1	0	0	0	SLR1-BP	S locus-related glycoprotein 1 binding pollen
pfam00759	22	10	11	17	26	9	5	7	Glyco_hydro_9	Glycosyl hydrolase family 9
pfam09770	20	5	9	22	12	24	2	12	PAT1	Topoisomerase II-associated protein PAT1
pfam03018	19	19	21	30	49	17	2	9	Dirigent	Dirigent-like protein
pfam00188	18	9	12	10	29	9	5	5	CAP	Cysteine-rich secretory protein family
pfam08263	16	30	11	10	37	12	6	5	LRRNT_2	Leucine rich repeat N-terminal domain
pfam14368	15	25	10	15	24	13	7	3	LTP_2	Probable lipid transfer

Continuing Table 5

	Dicots			Monocots			Mosses			
Pfam ID	*At*	*Pt*	*Sl*	*Bd*	*Osj*	*Zm*	*Pp*	*Sm*	Pfam name	Pfam discription
pfam00314	15	22	15	25	44	24	3	9	Thaumatin	Thaumatin family
pfam00067	12	31	20	31	74	13	6	21	p450	Cytochrome P450
pfam02298	12	20	14	30	36	28	9	9	Cu_bind_like	Plastocyanin-like domain
pfam04398	12	15	9	17	23	14	2	1	DUF538	Protein of unknown function
pfam02469	10	16	10	14	20	15	1	0	Fasciclin	Fasciclin domain
pfam01453	7	30	9	3	5	2	1	20	B_lectin	D-mannose binding lectin
pfam00197	7	22	15	2	3	0	0	0	Kunitz_legume	Trypsin and protease inhibitor
pfam00069	6	22	11	11	47	3	1	4	Pkinase	Protein kinase domain
pfam07714	6	19	4	3	24	4	0	1	Pkinase_Tyr	Protein tyrosine kinase
pfam00251	6	5	4	7	26	5	0	1	Glyco_hydro_32N	Glycosyl hydrolases family 32
pfam13947	4	23	2	6	32	7	0	0	GUB_WAK_bind	Wall-associated receptor kinase
pfam00704	2	14	8	13	31	10	1	10	Glyco_hydro_18	Glycosyl hydrolases family 18
pfam01559	0	0	0	0	0	30	0	0	Zein	Zein seed storage protein
pfam13352	0	0	0	0	0	0	61	0	DUF4100	Protein of unknown function (DUF4100)

Note: *At*: *Arabidopsis thaliana*; *Pt*: *Populus trichocarpa*; *Sl*: *Solanum lycopersicum*; Monocots - *Bd*: *Brachypodium distachyon*; *Osj*: *Oryza sativa (subsp. japonica)*; *Zm*: *Zea mays*. Mosses - Physcomitrella patens (subsp. patens); *Sm*: *Selaginella moellendorffii*. A complete list is in Supplementary Table 1

The functionalities of secreted proteins were further analyzed using rpsBLAST to search against Pfam in the Conserved Domain Database (CDD) (Marchler-Bauer et al., 2009). The results of Pfam analysis for a species having 20 or more members in a Pfam were summarized in Table 5. A complete list of Pfams can be found in Supplementary Table 1. The detailed analysis of molecular functions in secretomes searching Pfam revealed the difference in protein families among different species, including both variations in the number of members in a given Pfam and species specific Pfams (Table 5). Noticeably there were twice as many secreted peroxidase proteins in rice compared to Arabidopsis (Table 5). Plant peroxidases have multiple tissue-specific functions e.g., removal of hydrogen peroxide from chloroplasts and cytosol, oxidation of toxic compounds, biosynthesis of the cell wall, and defense responses towards wounding (Sottomayor and Barceló, 2004). The glycosyl hydrolases are suggested to have valuable applications in modifying plant cell wall architecture and in the development and characterization of new bioenergy and feedstocks (Lopez-Casado et al., 2008). The rice secretome consists of 31 members of Glyco-hydro-18 (GH18) and 26 of GH32N while only two GH18 and 6 GH32N were identified in the Arabidopsis secretome. We also observed a number of Pfams having more members in rice than in other species. These Pfams include dirigen-like protein, multicopper oxidase, pollen allergen, cytochrome P450, etc (Table 5). It should be noted that these predicted secreted cytochrome P450 proteins most likely are false positives as there is no secreted cytochrome P450 protein reported with experimental evidence in plants. Wen et al. (2007) reported a cytochrome P450 presented in the pea root cap secretome. However, its presence might represent leakage that occurs during the cell separation process. In general, moss species have fewer secreted proteins as well as a smaller member number in a given Pfam due to their small genomes. However, we noted that the lycophyte model organism *Selaginella moellendorffii* has 20 members of D-mannose binding lectin family, while other plant species have less than 10 members in this Pfam, except *Populus trichocarpa*, which has 30 members. Species-specific secreted proteins are also observed, such as corn, which has 30 members of Zein seed storage protein and *Physcomitrella patens* (subsp. patens), which has 61 members of a protein with an unknown function (DUF4100).

4 Discussion

We constructed the PlantSecKB to provide a resource for the plant research community. As the subcellular location(s) of a given protein curated by UniProtKB

or by us were considered first in assigning a subcellular location, these assignments are based on traceable literature with experimental evidence, and thus fairly reliable. However, the subcellular locations assigned based on the computational prediction will depend on the accuracy of the tools used. We have evaluated the prediction accuracy of the methods we used in this study and compared it with the accuracies of other methods (Table 1) (Min, 2010; Meinken and Min, 2012). We concluded the prediction of secreted proteins is relatively reliable. However, false positives and false negatives certainly exist. For example, a number of P450 enzymes were predicted to be secreted proteins, which are most likely false positives.

We also predicted other subcellular locations including mitochondrial, chloroplast, vacuole, nucleus, and others based on the predictions of TargetP and WoLF PSORT. Our evaluation on the prediction accuracies of these subcellular locations revealed that the accuracies of the tools we used, even though they are best among available tools, are still not satisfactory due to relatively low prediction sensitivities for these subcellular locations (Table 1) (Meinken and Min, 2013). With the exception of mitochondrial and cytosol proteins, however, the specificities for those subcellular locations including chloroplast, ER, Golgi apparatus, nucleus, plasma membrane, vacuole and cytoskeleton are acceptable (>89%). Thus, proteins predicted in those subcellular locations are relatively reliable, though they still need to be cautiously examined with experiments. Recently, several new tools were developed including the Cell-PLoc servers (Chou and Shen, 2008), MultiLoc2 (Blum et al., 2009), and others (Meinken and Min, 2012). These tools and their related publications can be found at our website (http://proteomics.ysu.edu/tools/subcell.html) (Meinken and Min, 2012). As standalone tools are not available for some of them, such as Cell-PLoc, or some standalone tools are too slow for processing a large data set, such as MutliLoc2, we were not able to use them for our data processing. However, we suggest users utilize these tools to get a second prediction for proteins of interest as our experience showed that using multiple tools improves prediction specificity.

Based on several recent large-scale secretome studies in plants, non-classical, i.e. leadless secretory proteins (LSPs) were observed to account for more than 50% of the total identified secretome, supporting the existence of novel secretory mechanisms independent of the classical ER-Golgi secretory pathway (Agrawal et al., 2010 for review; Jung et al., 2008; Cheng and Williamson, 2010; Ding et al., 2012). Mammalian and bacterial LSPs have been collected and used to implement the prediction software, SecretomeP, for predicting these proteins (http://www.cbs.dtu.dk/services/SecretomeP/) (Bendtsen et al., 2004a). Because the tool has not been trained with plant-specific data and the accuracy for predicting plant LSPs could not be evaluated, we did not include this tool in our data processing.

The PlantSecKB strives to serve as a portal for plant researchers to search plant protein subcellular locations with an emphasis on secreted proteins. The EST sub-database is expected to facilitate EST data mining for secreted proteins from expressed data, which is particularly useful for plant species not completely sequenced or having only a limited number of cDNA sequences. The collection and curation of secreted plant proteins, particularly LSPs, from literature with experimental evidence requires continuous efforts from the plant research community. We have implemented a curation tool accessible through PlantSecKB for the community to manually curate subcellular locations of plant proteins having experimental evidence. The utility described in PlantSecKB, together with our recently implemented Fungal Secretome KnowledgeBase (FunSecKB) (Lum and Min, 2011b), is anticipated to provide a search, download, and curation system that will help the plant community to further understand secretome biology. It can also be used to explore various potential applications and their interactions of plant and fungal secreted proteins for plant pathogen control and breeding for stress resistant varieties (Kim et al., 2009).

Authors' contributions

GL and JM implemented the database, JO and SF manually curated secreted proteins, XJM conceived of the study, designed the procedure of data processing.

XJM, JM and GL analyzed the data and prepared the manuscript. All authors read and approved the final manuscript.

Acknowledgments

The work is funded by the Ohio Plant Biotechnology Consortium [grant 2011-001] (through the Ohio State University, Ohio Agricultural Research and Development Center), and Youngstown State University (YSU) Research Council [grant 2010-2011 and 12-11]. The work is also supported by a YSU Research Professorship and the College of Science, Technology, Engineering, and Mathematics Dean's reassigned time for research to XJM. JM was supported with a graduate research assistantship by the Center for Applied Chemical Biology at YSU.

References

Agrawal G.K., Jwa N.S., Lebrun M.H., Job D., and Rakwal R., 2010, Plant secretome: unlocking secrets of the secreted proteins, Proteomics, 10: 799-827
http://dx.doi.org/10.1002/pmic.200900514

Alexandersson E., Ali A., Resjö S., and Andreasson E., 2013, Plant secretome proteomics, Front. Plant Sci., 4: 9
http://dx.doi.org/10.3389/fpls.2013.00009

Bendtsen J.D., Jensen L.J., Blom N., von Heijne G., and Brunak S., 2004a, Feature based prediction of non-classical and leaderless protein secretion, Protein Eng. Des. Sel., 17: 349-356
http://dx.doi.org/10.1093/protein/gzh037

Bendtsen J.D., Nielsen H., von Heijne G., and Brunak S., 2004b, Improved prediction of signal peptides: SignalP 3.0, J. Mol. Biol., 340: 783-795
http://dx.doi.org/10.1016/j.jmb.2004.05.028

Blum T., Briesemeister S., and Kohlbacher O., 2009, MultiLoc2: integrating phylogeny and Gene Ontology terms improves subcellular protein localization prediction, BMC Bioinformatics, 10: 274
http://dx.doi.org/10.1186/1471-2105-10-274

Borner G.H., Lilley K.S., Stevens T.J., and Dupree P., 2003, Identification of glycosylphosphatidylinositol-anchored proteins in Arabidopsis. A proteomic and genomic analysis, Plant Physiol., 132: 568-577
http://dx.doi.org/10.1104/pp.103.021170

Borner G.H., Sherrier D.J., Stevens T.J., Arkin I.T., and Dupree P., 2002, Prediction of glycosylphosphatidylinositol-anchored proteins in Arabidopsis. A genomic analysis, Plant Physiol., 129: 486-499
http://dx.doi.org/10.1104/pp.010884

Boudart G., Minic Z., Albenne C., Canut H., Jamet E., and Pont-Lezica R., 2007, Cell wall proteome, In: Samaj S., and Thelen J. (eds.), Plant Proteomics, Springer, pp.169-185

Bouws H., Wattenberg A., and Zorn H., 2008, Fungal secretomes-nature's toolbox for white biotechnology, Appl. Microbiol. Biotechnol., 80: 381-388
http://dx.doi.org/10.1007/s00253-008-1572-5

Caccia D., Dugo M., Callari M., and Bongarzone I., 2013, Bioinformatics tools for secretome analysis, Biochim. Biophys. Acta., S1570-9639

Chen X.Y., Kim S.T., Cho W.K., Rim Y., Kim S., Kim S.W., Kang K.Y., Park Z.Y., and Kim J.Y., 2009, Proteomics of weakly bound cell wall proteins in rice calli, J. Plant Physiol., 166: 675-685
http://dx.doi.org/10.1016/j.jplph.2008.09.010

Cheng F.Y., and Williamson J.D., 2010, Is there leaderless protein secretion in plants? Plant Signal Behav., 5: 129-131
http://dx.doi.org/10.4161/psb.5.2.10304

Cho W.K., and Kim J.Y., 2009, Integrated analyses of the rice secretome, Plant Signal Behav., 4: 345-347
http://dx.doi.org/10.4161/psb.4.4.8198

Cho W.K., Chen X.Y., Chu H., Rim Y., Kim S., Kim S.T., Kim S.W., Park Z.Y., and Kim J.Y., 2009, Proteomic analysis of the secretome of rice calli, Physiol. Plant, 135: 331-341
http://dx.doi.org/10.1111/j.1399-3054.2008.01198.x

Chou K.C., and Shen H.B., 2008, Cell-PLoc: a package of Web servers for predicting subcellular localization of proteins in various organisms, Nat. protoc., 3(2): 153-162
http://dx.doi.org/10.1038/nprot.2007.494

de Castro E., Sigrist C.J., Gattiker A., Bulliard V., Langendijk-Genevaux P.S., Gasteiger E., Bairoch A., and Hulo N., 2006, ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins, Nucleic Acids Res., 34(Web Server issue): W362-365

De-la-Peña C., Badri D.V., Lei Z., Watson B.S., Brandão M.M., Silva-Filho M.C., Sumner L.W., and Vivanco J.M., 2010, Root secretion of defense-related proteins is development-dependent and correlated with flowering time, J. Biol. Chem., 285: 30654-30665
http://dx.doi.org/10.1074/jbc.M110.119040

Ding Y., Wang J., Wang J., Stierhof Y.D., Robinson D.G., and Jiang L., 2012, Unconventional protein secretion, Trends Plant Sci., 7: 606-615
http://dx.doi.org/10.1016/j.tplants.2012.06.004

Duvick J., Fu A., Muppirala U., Sabharwal M., Wilkerson M.D., Lawrence C.J., Lushbough C., and Brendel V., 2008,

PlantGDB: a resource for comparative plant genomics, Nucl. Acids Res., 36: D959-965
http://dx.doi.org/10.1093/nar/gkm1041

Emanuelsson O., Brunak S., von Heijne G., and Nielsen H., 2007, Locating proteins in the cell using TargetP, SignalP and related tools, Nat. Protoc., 2: 953-971
http://dx.doi.org/10.1038/nprot.2007.131

Emanuelsson O., Nielsen H., Brunak S., and von Heijne G., 2000, Predicting subcellular localization of proteins based on their N-terminal amino acid sequence, J. Mol. Biol., 300: 1005-1016
http://dx.doi.org/10.1006/jmbi.2000.3903

Finnie C., Andersen B., Shahpiri A., and Svensson B., 2011, Proteomes of the barley aleurone layer: A model system for plant signalling and protein secretion, Proteomics, 11: 1595-1605
http://dx.doi.org/10.1002/pmic.201000656

Foresti O., and Denecke J., 2008, Intermediate organelles of the plant secretory pathway: identity and function, Traffic, 9: 1599-1612
http://dx.doi.org/10.1111/j.1600-0854.2008.00791.x

Gillmor C.S., Lukowitz W., Brininstool G., Sedbrook J.C., Hamann T., Poindexter P., and Somerville C., 2005, Glycosylphosphatidylinositol-anchored proteins are required for cell wall synthesis and morphogenesis in Arabidopsis, Plant Cell, 17:1128-1140
http://dx.doi.org/10.1105/tpc.105.031815

Greenbaum D., Luscombe N.M., Jansen R., Qian J., and Gerstein M., 2001, Interrelating different types of genomic data, from proteome to secretome: coming in on function, Genome Res., 11: 1463-1468
http://dx.doi.org/10.1101/gr.207401

Hathout Y., 2007, Approaches to the study of the cell secretome, Expert Rev. Proteomics, 4: 239-248
http://dx.doi.org/10.1586/14789450.4.2.239

Horton P., Park K.J., Obayashi T., Fujita N., Harada H., Adams-Collier C.J., and Nakai K., 2007, WoLF PSORT: protein localization predictor. Nucleic acids res., 35(Web Server issue): W585-587

Isaacson T., and Rose J.K.C., 2006, The plant cell wall proteome, or secretome, In Plant Proteomics, Annual Plant Reviews Series, edited by Finnie C., Blackwell Publishing, 28:185-209

Jamet E., Albenne C., Boudart G., Irshad M., Canut H., and Pont-Lezica R, 2008, Recent advances in plant cell wall proteomics, Proteomics, 8: 893-908
http://dx.doi.org/10.1002/pmic.200700938

Jones R.L., and Robinson D.G., 1989, Protein Secretion in Plants, Tansley Review No. 17, New Phytologist, 111: 567-597
http://dx.doi.org/10.1111/j.1469-8137.1989.tb02352.x

Jung Y.H., Jeong S.H., Kim S.H., Singh R., Lee J.E., Cho Y.S., Agrawal G.K., Rakwal R., and Jwa N.S., 2008, Systematic secretome analyses of rice leaf and seed callus suspension-cultured cells: workflow development and establishment of high-density two-dimensional gel reference maps, J. Proteome Res., 7: 5187-5210
http://dx.doi.org/10.1021/pr8005149

Käll L., Krogh A., and Sonnhammer E.L.L., 2007, Advantages of combined transmembrane topology and signal peptide prediction--the Phobius web server, Nucleic acids res., 35(Web Server issue): W429-432

Kamoun S., 2009, The Secretome of Plant-Associated Fungi and Oomycetes, In: Deising V.H. (ed.), Plant Relationships, 2nd Edition, The Mycota, Springer-Verlag, Berlin Heidelberg, pp 173-180

Kim S.T., Kang Y.H., Wang Y., Wu J., Park Z.Y., Rakwal R., Agrawal G.K., Lee S.Y., and Kang K.Y., 2009, Secretome analysis of differentially induced proteins in rice suspension-cultured cells triggered by rice blast fungus and elicitor, Proteomics, 9: 1302-1313
http://dx.doi.org/10.1002/pmic.200800589

Krause C., Richter S., Knöll C., and Jürgens G., 2013, Plant secretome - From cellular process to biological activity, Biochim. Biophys. Acta, 1834(11): 2429-2441

Krogh A., Larsson B., von Heijne G., and Sonnhammer E.L.L., 2001, Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes, J. Mol. Biol., 305: 567-580
http://dx.doi.org/10.1006/jmbi.2000.4315

Kusumawati L., Imin N., and Djordjevic M.A., 2008, Characterization of the secretome of suspension cultures of Medicago species reveals proteins important for defense and development, J. Proteome Res., 7: 4508-4520
http://dx.doi.org/10.1021/pr800291z

Lopez-Casado G., Urbanowicz B.R., Damasceno C.M.B., and Rose J.K.C., 2008, Plant glycosyl hydrolases and biofuels: a natural marriage, Current Opinion Plant Biol., 11: 329-337
http://dx.doi.org/10.1016/j.pbi.2008.02.010

Lum G., Vanburen R., Ming R., Min X.J., 2013, Secretome prediction and analysis in sacred lotus (*Nelumbo nucifera Gaertn.*), Tropical Plant Biol., 6:131-137
http://dx.doi.org/10.1007/s12042-013-9121-5

Lum G., and Min X.J., 2013, Bioinformatic protocols and the knowledge-base for secretomes in fungi, In: Gupta V.K., Tuohy M.G., Ayyachamy M., Turner K.M. and O'Donovan A. (eds.), Laboratory Protocols in Fungal

Biology: Current Methods in Fungal Biology, Springer, pp 545-557
http://dx.doi.org/10.1007/978-1-4614-2356-0_54

Lum G., and Min X.J., 2011a, Plant secretomes: Current status and future perspectives, Plant Omics J., 4: 114-119

Lum G., and Min X.J., 2011b, FunSecKB: the fungal secretome knowledgebase, Database - J. Biol. Databases Curation, Vol. 2011
http://dx.doi.org/10.1093/database/bar001

Marchler-Bauer A., Lu S., Anderson J.B., Chitsaz F., Derbyshire M.K., DeWeese-Scott C., Fong J.H., Geer L.Y., Geer R.C., Gonzales N.R., Gwadz M., Hurwitz D.I., Jackson J.D., Ke Z., Lanczycki C.J., Lu F., Marchler G.H., Mullokandov M., Omelchenko M.V., Robertson C.L., Song J.S., Thanki N., Yamashita R.A., Zhang D., Zhang N., Zheng C., and Bryant S.H., 2011, CDD: a Conserved Domain Database for the functional annotation of proteins, Nucleic Acids Res., 39(Database issue): D225-229
http://dx.doi.org/10.1093/nar/gkq1189

McCarthy F.M., Wang N., Magee G.B., Nanduri B., Lawrence M.L., Camon E.B., Barrell D.G.,Hill D.P., Dolan M.E., Williams W.P., Luthe D.S., Bridges S.M., and Burgess S.C., 2006, AgBase: a functional genomics resource for agriculture, BMC Genomics, 7: 229
http://dx.doi.org/10.1186/1471-2164-7-229

Meinken J., and Min X.J., 2012, Computational prediction of protein subcellular locations in eukaryotes: an experience report. Comput. Mole. Biol., 2(1): 1-7

Melhem H., Min X.J., and Butler G., 2013, The impact of SignalP 4.0 on the prediction of secreted proteins, 2013 IEEE Symposium Series on Computational Intelligence (IIEEE SSCI 2013): The 10th annual IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, Singapore, pp.16-22

Min X.J., 2010, Evaluation of computational methods for secreted protein prediction in different eukaryotes, J. Proteomics Bioinform., 3: 143-147

Min X.J., Butler G., Storms R., and Tsang A., 2005a, OrfPredictor: predicting protein-coding regions in EST-derived sequences, Nucleic Acids Res., 33: W677-680
http://dx.doi.org/10.1093/nar/gki394

Min X.J., Butler G., Storms R., and Tsang A., 2005b, TargetIdentifier: a web server for identifying full-length cDNAs from EST sequences, Nucleic Acids Res., 33: W669-672
http://dx.doi.org/10.1093/nar/gki436

Ming R., Vanburen R., Liu Y., Yang M., Han Y., Li L.T., Zhang Q., Kim M.J., Schatz M.C.,Campbell M., Li J., Bowers J.E., Tang H., Lyons E., Ferguson A.A., Narzisi G., Nelson D.R., Blaby-Haas C.E., Gschwend A.R., Jiao Y., Der J.P., Zeng F., Han J., Min X.J., Hudson K.A.,Singh R., Grennan A.K., Karpowicz S.J., Watling J.R., Ito K., Robinson S.A., Hudson M.E., Yu Q., Mockler T.C., Carroll A., Zheng Y., Sunkar R., Jia R., Chen N., Arro J., Wai C.M., Wafula E., Spence A., Han Y., Xu L., Zhang J., Peery R., Haus M.J., Xiong W., Walsh J.A., Wu J., Wang M.L., Zhu Y.J., Paull R.E., Britt A.B., Du C., Downie S.R., Schuler M.A., Michael T.P., Long S.P., Ort D.R., Schopf J.W., Gang D.R., Jiang N., Yandell M., Depamphilis C.W., Merchant S.S., Paterson A.H., Buchanan B.B., Li S., Shen-Miller J., 2013, Genome of the long-living sacred lotus (*Nelumbo nucifera Gaertn.*), Genome Biol., 14(5): R41
http://dx.doi.org/10.1186/gb-2013-14-5-r41

Ngara R., and Ndimba B.K., 2011, Mapping and characterization of the sorghum cell suspension culture secretome, African J. Biotechnol., 10: 253-266

Oh I.S., Park A.R., Bae M.S., Kwon S.J., Kim Y.S., Lee J.E., Kang N.Y., Lee S., Cheong H., and Park O.K., 2005, Secretome analysis reveals an Arabidopsis lipase involved in defense against *Alternaria brassicicola*, Plant Cell, 17: 2832-2847
http://dx.doi.org/10.1105/tpc.105.034819

Petersen T.N., Brunak S., von Heijne G., and Nielsen H., 2011, SignalP 4.0: discriminating signal peptides from transmembrane regions, Nature Methods, 8: 785-786
http://dx.doi.org/10.1038/nmeth.1701

Poisson G., Chauve C., Chen X., and Bergeron A., 2007, FragAnchor a large scale all Eukaryota predictor of Glycosylphosphatidylinositol-anchor in protein sequences by qualitative scoring, Genomics Proteomics Bioinform., 5: 121-130
http://dx.doi.org/10.1016/S1672-0229(07)60022-9

Ranki H., and Sopanen T., 1984, Secretion of alpha-amylase by the aleurone layer and the scutellum of germinating barley grain, Plant Physiol., 75: 710-715
http://dx.doi.org/10.1104/pp.75.3.710

Rose J.K., and Lee S.J., 2010, Straying off the highway: trafficking of secreted plant proteins and complexity in the plant cell wall proteome, Plant Physiol., 153: 433-436
http://dx.doi.org/10.1104/pp.110.154872

Shinano T., Komatsu S., Yoshimura T., Tokutake S., Kong F.J., Watanabe T., Wasaki J., Osaki M., 2011, Proteomic analysis of secreted proteins from aseptically grown rice, Phytochemistry, 72: 312-320
http://dx.doi.org/10.1016/j.phytochem.2010.12.006

Sigrist, C.J.A., Cerutti, L., de Casro, E., Langendijk-Genevaux, P.S., Bulliard, V., Bairoch, A., and Hulo N., 2010, PROSITE, a protein domain database for functional characterization and annotation, Nucleic Acids Res., 38: 161-166
http://dx.doi.org/10.1093/nar/gkp885

Simpson C., Thomas C., Findlay K., Bayer E., and Maule A.J., 2009, An Arabidopsis GPI-anchor plasmodesmal neck protein with callose binding activity and potential to regulate cell-to-cell trafficking, Plant Cell, 21: 581-594
http://dx.doi.org/10.1105/tpc.108.060145

Sottomayor M., and Barceló A.R., 2004, Plant peroxidases and phytochemistry – foreword, Phytochemistry Rev., 3: 1-2
http://dx.doi.org/10.1023/B:PHYT.0000047819.21396.15

Tjalsma H., Bolhuis A., Jongbloed J.D., Bron S., and van Dijl J.M., 2000, Signal peptide-dependent protein transport in Bacillus subtilis: a genome-based survey of the secretome, Microbiol. Mol. Biol. Rev., 64: 515-547
http://dx.doi.org/10.1128/MMBR.64.3.515-547.2000

von Heijne G., 1990, The signal peptide, J. Membr. Biol., 115: 195-201
http://dx.doi.org/10.1007/BF01868635

Wen F., VanEtten H.D., Tsaprailis G., and Hawes M.C., 2007, Extracellular proteins in pea root tip and border cell exudates, Plant Physiol., 143: 773-783
http://dx.doi.org/10.1104/pp.106.091637

Werck-Reichhart D., and Feyereisen R., 2000, Cytochromes P450: a success story, Genome Biol., 1: REVIEWS3003

Zhang L., Tian L.H., Zhao J.F., Song Y., Zhang C.J., and Guo Y., 2009, Identification of an apoplastic protein involved in the initial phase of salt stress response in rice root by two-dimensional electrophoresis, Plant Physiol., 149: 916-928
http://dx.doi.org/10.1104/pp.108.131144

Table 2 Summary of subcellular proteomes in different plant species in PlantSecKB

	Total	Sec	Mt		Ch		Cyt	Ctk	Gol	Per	Nuc	Pla mem	Vac	GPI anc
			mem	non-mem	mem	non-mem								
Green algae														
Chlamydomonas reinhardtii	15120	545	435	3422	256	1608	2688	71	14	52	1641	487	131	27
Volvox carteri	14825	569	470	3019	200	1445	2478	46	15	59	2120	490	109	36
Micromonas pusilla	10311	155	242	1708	268	1854	2005	36	11	21	1557	231	83	6
Micromonas sp.	10133	184	248	1551	284	1494	2204	42	11	34	1595	324	86	38
Chlorella variabilis	9836	411	286	2018	136	910	1873	19	16	47	1213	361	80	33
Coccomyxa subellipsoidea	9799	426	210	1551	101	786	2121	36	23	27	1596	452	97	11
Ostreococcus tauri	8029	71	274	1979	186	1234	1376	12	8	4	1037	212	42	9
Bathycoccus prasinos	7886	145	172	918	304	1646	1097	34	16	13	1877	310	54	15
Ostreococcus lucimarinus	7404	50	235	1605	72	573	1780	40	8	12	923	200	69	8
Monocots														
Oryza sativa subsp. japonica	99948	5027	1707	17419	1329	14569	15442	545	213	144	19654	3226	736	366
Zea mays	62866	3888	1161	10333	1106	9632	9812	357	135	105	10910	1899	540	373
Oryza sativa subsp. indica	40429	2646	772	6199	661	5391	6686	228	71	66	7274	1714	405	232
Setaria italica	39296	2436	785	6748	699	5692	6019	198	62	64	6784	1645	342	225
Sorghum bicolor	33979	2096	625	5211	549	4681	5876	170	51	70	6366	1388	297	242
Oryza brachyantha	32339	1969	615	5406	400	4027	5075	183	40	54	5785	1406	308	127
Oryza glaberrima	32094	2151	674	5132	536	4689	4890	187	55	53	5625	1375	319	216
Brachypodium distachyon	30180	2204	627	4416	594	4339	4502	216	59	52	5768	1552	287	208
Hordeum vulgare	21743	1584	538	3397	443	3722	3101	111	37	43	3357	1086	192	186
Dicots														
Glycine max	74114	4369	1004	7750	1296	8956	12621	548	235	107	17228	4270	799	378
Medicago truncatula	56371	3120	546	6589	618	5569	11747	367	147	132	11791	2009	458	165
Vitis vinifera	54268	2429	678	6310	718	5586	10743	383	106	95	12360	2447	389	148
Arabidopsis thaliana	53847	3696	782	5470	1051	6225	9855	494	454	177	14076	2516	646	251
Populus trichocarpa	45325	2375	512	4760	609	4531	9349	374	138	79	10030	2100	540	179

Continuing Table 2

	Total	Sec	Mt		Ch		Cyt	Ctk	Gol	Per	Nuc	Pla mem	Vac	GPI anc
			mem	non-mem	mem	Non-mem								
Solanum lycopersicum	36341	2209	352	3768	523	3817	6816	274	81	76	8264	1809	376	134
Arabidopsis lyrata	32797	2447	383	3404	527	4159	5918	216	90	42	7608	1695	343	192
Ricinus communis	31471	1848	359	4133	480	3648	5911	217	74	51	6374	1424	322	110
Nelumbo nucifera	26899	1313	440	3696	585	3611	4446	196	59	40	5350	1352	261	98
Lotus japonicus	8674	555	90	1041	172	1071	1701	64	27	17	1595	269	129	47
Mosses														
Physcomitrella patens	34939	781	413	4641	286	3519	9058	287	79	66	7542	1080	218	35
Selaginella moellendorffii	33294	1749	600	4718	317	2389	7694	267	85	57	5837	1510	287	88
Conifer														
Picea sitchensis	11307	578	137	1487	221	1426	2269	85	25	24	2186	356	129	64
Total for all Species	1415921	66063	25845	173076	47635	237882	252258	9931	2755	2669	241354	54497	13114	5109

Note: Sec: secretome; Mt: mitochondrial; mem: membrane; non-mem: non-membrane; Ch: chloroplast; Cyt: cytosol; Ctk: cytoskeleton; Gol: Golgi apparatus; Per: peroxisome; Vac: vacuole; Pla mem: plasma membrane; GPI anc: glycosylphosphatidyinositol anchored