Plant Omics Journal POJ

POJ 4(2):114-119 (2011) ISSN:1836-3644

Research Note

Plant Secretomes: Current status and future perspectives

Gengkon Lum[1], Xiang Jia Min[2*]

[1]Department of Computer Science and Information Systems, Youngstown State University, OH 44555, USA
[2]Center for Applied Chemical Biology, Department of Biological Sciences, Youngstown State University, OH 44555, USA

***Corresponding author: xmin@ysu.edu**

Abstract

Plant secreted proteins have biological functions which are important in the formation of cell walls, cellular communication, and defense against pathogens. We analyzed 1704 secreted proteins from a total of 22513 plant proteins, all of which were manually curated and annotated in the UniProt database. Of the secreted plant proteins analyzed, 55% and 13% are curated from *Arabidopsis thaliana* and rice, respectively. Gene ontology cellular components analysis revealed that 84% of them are located in the extracellular region, cell wall, or extracellular space. Molecular functional domain analysis showed that 33% had hydrolase activity and 29% had binding activity. Signal peptide analysis revealed that 97.5% of secreted proteins had signal peptides. The information is anticipated to be used to computationally identify more secreted proteins in plants and to construct a plant secretome knowledge database.

Keywords: plant, secreted protein, secretome, signal peptide.
Abbreviations: GO: Gene Ontology, LSP: leaderless secretory protein.

Introduction

Plant secreted proteins play important biological roles in cell wall structure, cellular communication, and the host-pathogen relationships (Isaacson and Rose, 2006; Kamoun, 2009). One well studied system was germinating barley seed in which α-amylase was found to be synthesized in the aleurone layer and secreted into the endosperm to break down starch (Ranki and Sopanen, 1984; Jones and Robinson, 1989 for review). More 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 for review). The term secretome was first used to describe genome-wide study of the signal peptide-dependent secreted proteins and the protein secretion machineries in *Bacillus subtilis*, a Gram-positive bacterium (Tjalsma et al., 2000). Though it is still occasionally used to include the set of proteins involved in the secretory pathway (Simpson et al., 2007), however, the term is more often limited to include only the secreted proteins (Greenbaum et al., 2001; Hathout, 2007; Bouws et al., 2008). Recently Agrawal et al. (2010) comprehensively reviewed the state of progress in plant secretomics research, including experimental systems and techniques for identification of secreted proteins in plants. A revised secretome definition was proposed as "the global group of secreted proteins into the extracellular space by a cell, tissue, organ or organism at any given time and conditions through known and unknown secretory mechanisms involving constitutive and regulated secretory organells" (Agrawal et al., 2010). Thus in this work, plant secretomes refer to all proteins which are secreted into the extracellular regions or extracellular space, i. e., outside of the plasma membrane, of plant cells or tissues. With the improvement of sequencing technology and the reduced cost of sequencing, the genomes of more plant species are being completely sequenced. Currently there are 24 land plants having complete or draft genome sequences available and 72 land plant species with genome sequencing in progress (http://www.ncbi.nlm.nih.gov/genomes/static/gpstat.html). The ability to predict the protein coding genes and the subcellular locations of these encoded proteins is essential for the functional annotation of the genomes. In an attempt to better predict and analyze all secreted proteins, i. e. secretomes, from plants, we analyzed all the plant secreted proteins so far manually curated and annotated in the UniProt database (http://www.uniprot.org/) (The UniProt Consortium, 2010). The information obtained through this

Table 1. Plant species distribution of curated secreted proteins in UniProt/Swiss-Prot database.

Species	Number of proteins
Arabidopsis thaliana (Mouse-ear cress)	941
Oryza sativa subsp. japonica (Rice)	226
Solanum lycopersicum (Tomato) (Lycopersicon esculentum)	37
Nicotiana tabacum (Common tobacco).	28
Hordeum vulgare (Barley)	27
Triticum aestivum (Wheat)	25
Zea mays (Maize)	21
Oryza sativa subsp. indica (Rice)	16
Capsicum annuum (Bell pepper)	12
Betula verrucosa (White birch) (Betula pendula)	11
Cycas revoluta (Sago palm)	10
Phaseolus vulgaris (Kidney bean) (French bean)	10
Solanum tuberosum (Potato)	10
Others (153 species)	330
Total	1704

analysis is anticipated to be useful in developing methods to accurately predict plant secretomes for the construction of a plant secretome database.

Data and methods

UniProtKB consists of two datasets, UniProtKB/Swiss-Prot and UniProtKB/TrEMBL (http://www.uniprot.org/). UniProtKB/Swiss-Prot contains manually annotated non-redundant protein sequences with information extracted from literature of experimental results and curator-evaluated computational analysis (The UniProt Consortium, 2010). UniProtKB/TrEMBL contains protein sequences associated with computationally generated annotation and large-scale functional characterization. All the entries belonging to kingdom *Viridiplantae* having a subcellular location annotated as "secreted" in the UniProtKB/Swiss-Prot dataset were retrieved (http://www.uniprot.org/downloads; Release-2010_09). The dataset consisted of 1704 secreted proteins within a total of 22513 proteins. The species distribution of secreted proteins was analyzed by retrieving the species information associated with each secreted protein. Gene ontology (GO) information was retrieved from the dataset and analyzed using GO SlimViewer with plant specific GO terms (McCarthy et al., 2006). The protein families and functional domains of plant secreted proteins were analyzed using rpsBLAST and searched against Pfam and the Conserved Domain Database (Marchler-Bauer et al., 2009). The existence of a signal peptide was predicted using three predictors including SignalP (version 3.0, http://www.cbs.dtu.dk/services/SignalP/) (Bendtsen et al., 2004b), Phobius (http://phobius.binf.ku.dk/) (Käll et al., 2004) and TargetP (http://www.cbs.dtu.dk/services/TargetP/) (Emanuelsson et al., 2007). We chose these three predictors because they were widely used and were previously evaluated favorably (Min, 2010). TMHMM (http://www.cbs.dtu.dk/ services/TMHMM) was used to identify proteins having transmembrane domains (Emanuelsson et al., 2007). The default parameters were used for all the programs. For SignalP prediction, only entries that are predicted to have a "mostly likely cleavage site" by SignalP-NN algorithm and a "signal peptide" by SignalP-HMM algorithm are considered to be true signal peptide "positives" using the N-terminal 70 amino acids (Bendtsen et al., 2004b). For predicting membrane proteins using TMHMM, the entries having membrane domains not located within the N-terminus (the first 70 amino acids) were treated as real membrane proteins.

Results

Species distribution of curated secreted proteins in plants

A total of 1704 secreted proteins from 166 plant species have been manually curated so far in the UniProtKB/Swiss-Prot dataset (UniProt-Release-2010_09). However, it should be noted that 1340 of them were annotated with non-experimental qualifiers including "Potential" (439 proteins), "Probable" (119 proteins) or "By similarity" (792 proteins) in their subcellular locations (http://www.uniprot.org/manual/non_experimental_qualifiers). Thus, the sub-cellular locations of the majority of the curated secreted proteins have not yet been experimentally verified. Among the 166 plant species, 153 of them have less than 10 entries. Plant species having 10 or more entries are listed in Table 1. *A. thaliana*, as the most intensively studied model plant, has the highest number of curated secreted proteins at 941 and *O. sativa* (subsp. Japonica) (rice) has 226 secreted proteins curated in the database (Table 1).

GO analysis

The annotated GO IDs were retrieved for all plant secreted proteins. They were then mapped to top categories using GO SlimViewer (McCarthy et al., 2006). There are three main categories of GO terms including biological processes, cellular components, and molecular functions (http://www.geneontology.org/). As one protein may have multiple GO terms, the total of GO terms in each category is more than the total number of curated proteins. Overall the distribution patterns of each GO category of secreted proteins in the whole set and in *Arabidopsis* are similar (Table 2). GO analysis shows that secreted proteins play important roles in many biological processes including various cellular process, metabolic and catabolic process, responses to stress and biotic stimulus, cellular component organization, etc. Cellular component analysis showed that ~86% of the curated secreted proteins are located extracellularly including extracellular region, extracellular space, or cell wall. As these proteins were curated as secreted proteins, the reason for some to be found as other subcellular components was that some entries were annotated to have multiple subcellular locations and some entries were secreted but attached to the outside of the plasma membrane and counted as membrane components. GO analysis of molecular functions showed that about 38%

Table 2. Gene Ontology analysis of curated secreted proteins in UniProt/Swiss-Prot in all plants and Arabidopsis thaliana.

		All Plants		Arabidopsis thaliana	
Biological processes		Number of GO terms	%[a]	Number of GO terms	%[a]
GO:0009987	cellular process	705	20	383	17
GO:0006950	response to stress	675	19	419	19
GO:0009607	response to biotic stimulus	410	12	311	14
GO:0009056	catabolic process	388	11	228	10
GO:0016043	cellular component organization	385	11	208	9
GO:0005975	carbohydrate metabolic process	265	7	123	6
GO:0006629	lipid metabolic process	114	3	107	5
GO:0019538	protein metabolic process	71	2	56	3
GO:0000003	reproduction	64	2	32	1
	others	473	13	330	15
	Total	3550		2197	
Cellular components					
GO:0005576	extracellular region	1639	68	940	70
GO:0005618	cell wall	431	18	219	16
GO:0016020	membrane	174	7	97	7
GO:0005615	extracellular space	86	4	18	1
GO:0005773	vacuole	47	2	32	2
GO:0005737	cytoplasm	20	1	12	1
	others	40	2	25	2
	Total	2428		1343	
Molecular functions					
GO:0016787	hydrolase activity	653	38	427	49
GO:0005488	binding	583	34	260	30
GO:0003824	catalytic activity	234	14	100	11
GO:0030234	enzyme regulator activity	182	11	45	5
GO:0016740	transferase activity	52	3	37	4
	others	22	1	7	1
	Total	1726		876	

[a] The percentage (%) of each GO term subcategory relative to the total GO terms in each category

of the whole secreted set and 49% of *Arabidopsis* secreted proteins have hydrolase activity, about one third have binding activity, and 11- 14% have catalytic activities (Table 2).

Protein family and conserved domain analysis

Protein families and conserved domain analysis were carried out using rpsBLAST to search the Pfam database first and, for no hit proteins, to search the CDD domain database with a cutoff E-value of 1e-10. A total of 1196 secreted proteins were identified to have protein families or conserved domains. Those families or domains having 10 or more secreted proteins were listed in Table 3. The majority of protein families were enzymes involved in cell wall formation including hydrolases (divided into several families), cell wall structure proteins (pollen allergen family which include expansin subfamilies), enzymes involved in defense such as peroxidase, and the cupin family, which are plant seed storage proteins and germins.

Signal peptide and transmembrane domain prediction

After removing protein entries not starting with a methionine (M) (assumed to be partial), we examined the presence of signal peptides and transmembrane domains in the 1497 curated secreted full-length proteins (Table 4). The number of positives in Table 4 refers the proteins predicted to have a signal peptide by SignalP, Phobius, and TargetP, or not having a transmembrane domain by TMHMM. The results showed that >90% of the curated secreted proteins in plants consisting of a signal peptide sequence, that could be detected by one of the tools used. When two or three signal peptide predictors were used additively, >97% of secreted proteins were predicted to have a signal peptide. Thus, the presence of a signal peptide is a reliable indicator for secreted protein prediction and curation. Transmembrane domain analysis showed that a small number of curated secreted proteins (1.5%) also contained a transmembrane domain.

Discussion

UniProt annotation project has curated more than 1700 secreted proteins from plant species; with 68% of these from Arabidopsis and rice (Table 1). Among the curated secreted proteins, however, 79% with subcellular locations were not experimentally verified. GO analysis and functional domain analysis showed that plant secreted proteins play important roles in diverse biological processes (Table 2 and 3). Particularly they are involved in cell wall formation as enzymes such as hydrolases or structure proteins such as expansins (Table 3) (Boudart et al., 2007; LoPez-Casado et al., 2008; Sampedro and

Table 3. Functional domain analysis of plant secreted proteins.

Families/Domains	Proteins
pfam00141, peroxidase.	102
pfam00190, Cupin_1, Cupin.	88
pfam01357, Pollen_allerg_1, Pollen allergen.	88
cd01837, SGNH_plant_lipase_like, a plant specific subfamily of the SGNH-family of hydrolases, a diverse family of lipases and esterases.	74
pfam01095, Pectinesterase.	68
pfam00450, Peptidase_S10, Serine carboxypeptidase.	56
pfam07732, Cu-oxidase_3, Multicopper oxidase.	48
pfam01657, DUF26, Domain of unknown function DUF26.	43
pfam00759, Glyco_hydro_9, Glycosyl hydrolase family 9.	42
pfam00722, Glyco_hydro_16, Glycosyl hydrolases family 16.	41
pfam01301, Glyco_hydro_35, Glycosyl hydrolases family 35.	36
pfam00657, Lipase_GDSL, GDSL-like Lipase/Acylhydrolase.	31
pfam00149, Metallophos, Calcineurin-like phosphoesterase.	26
pfam00295, Glyco_hydro_28, Glycosyl hydrolases family 28.	26
cd00261, AAI_SS, Alpha-amylase inhibitors and seed storage protein subfamily.	25
pfam07333, SLR1-BP, S locus-related glycoprotein 1 binding pollen coat protein.	24
pfam00251, Glyco_hydro_32N, Glycosyl hydrolases family 32	21
pfam00304, Gamma-thionin, Gamma-thionin family.	20
pfam00332, Glyco_hydro_17, Glycosyl hydrolases family 17.	18
pfam00321, Thionin, Plant thionin.	17
pfam09265, Cytokin-bind, Cytokinin dehydrogenase 1, FAD and cytokinin binding.	17
COG3934, COG3934, Endo-beta-mannanase.	16
pfam03330, DPBB_1, Rare lipoprotein A (RlpA)-like double-psi beta-barrel.	15
pfam00182, Glyco_hydro_19, Chitinase class I.	12
pfam06404, PSK, Phytosulfokine precursor protein.	12
pfam00445, Ribonuclease_T2, Ribonuclease T2 family.	11
pfam00197, Kunitz_legume, Trypsin and protease inhibitor.	10
pfam00314, Thaumatin, Thaumatin family.	10
others (73 families)	199
Total	1196

Table 4. Prediction of signal peptides and transmembrane domains of curated plant secreted.

Tools	Number of Positives	Number of Negatives	Positives (%)
SignalP	1398	99	93.4
Phobius	1367	130	91.3
TargetP	1415	82	94.5
SignalP+TargetP[a]	1356	141	90.6
SignalP/TargetP[b]	1457	40	97.3
SignalP+Phobius+TargetP[a]	1297	200	86.6
SignalP/Phobius/TargetP[b]	1460	37	97.5
TMHMM[c]	1474	23	98.5

[a] The signal peptide is detected by all tools.
[b] The signal peptide is detected by any one of the tools.
[c] Entries not having a transmembrane domain are treated as positives.

Cosgrove, 2005). More than 100 secreted peroxidases were also curated, these enzymes 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). Using signal peptide predicting tools to examine the presence of signal peptides in the curated secreted proteins revealed that >97% of entries had a signal peptide (Table 4). Thus, the presence of signal peptide remains to be an effective indicator for identifying secreted proteins in plants. We have evaluated the accuracies of these prediction tools and proposed to combine multiple signal prediction tools with TMHMM and PS-Scan to predict potential signal peptide containing secreted proteins from predicted proteomes of completely sequenced plant genomes (Min, 2010). Transmembrane domain analysis showed that 1.5% of manually curated proteins were predicted to have transmembrane domains (Table 4). Whether these entries are real membrane proteins still remains to be investigated.

Recent studies, however, which used *in vitro* suspension cultured cells and *in planta* systems, identified a large number of leaderless secretory proteins (LSPs) in several plant species including Arabidopsis, rice, and *Medicago* species (Tran and Plaxton, 2008; Jung et al., 2008; Cho et al., 2009; Kusumawati et al., 2008; Agrawal et al., 2010). These LSPs can account for, on average, more than 50% of the total identified secretome, supporting the existence of a novel signal peptide independent secretory mechanism. These LSPs were mainly identified under biotic and abiotic stress conditions, suggesting their involvement in defense or stress responses. Clearly, these LSPs have not been curated in the UniProt annotation projects. Non-classical, signal peptide independent, secretion pathways may exist in all domains of organisms from bacteria to human. Mammalian and bacterial LSPs have been collected and used to implement the prediction software, SecretomeP, for predicting LSPs (Bendtsen et al., 2004a; Bendtsen et al., 2005) (http://www.cbs.dtu.dk/services/SecretomeP/). However, there is no plant-specific software tool available yet for predicting the LSPs. Collecting and curating these LSPs in plants will be essential in developing a method to predict potential LSPs in plants. In addition to the UniProtKB, a dedicated plant secretome database to curate all secreted plant secreted proteins, including the LSPs, will be a useful resource for the plant research community. In considering the important roles played by secreted proteins in plant defense and cell wall biosynthesis, the study of plant secretomes may lead to the breeding of plants more resistant to pathogens and stresses for use in food and bio-fuel production.

Acknowledgment

We thank Dr. Gary Walker for providing helpful comments on the paper and Jessica Orr for assistance in data collection. The work is supported by the Ohio Plant Biotechnology Consortium (grant 2011-001) (through the Ohio State University, Ohio Agricultural Research and Development Center), Youngstown State University (YSU) Research Council (grant 2010-2011 #12-11), YSU Research Professorship, and the College of Science, Technology, Engineering, and Mathematics Dean's reassigned time for research to XJM.

References

Agrawal GK, Jwa NS, Lebrun MH, Job D, Rakwal R (2010) Plant secretome: unlocking secrets of the secreted proteins. Proteomics 10:799-827

Bendtsen JD, Jensen LJ, Blom N, von Heijne G, Brunak S (2004a) Feature based prediction of non-classical and leaderless protein secretion. Protein Eng Des Sel 17:349-356

Bendtsen JD, Kiemer L, Fausbøll A, Brunak S (2005) Non-classical protein secretion in bacteria. BMC Microbiol 5:58

Bendtsen JD, Nielsen H, von Heijne G, Brunak S (2004b) Improved prediction of signal peptides: SignalP 3.0. J Mol Biol 340, 783-795

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

Bouws H, Wattenberg A, Zorn H (2008) Fungal secretomes-nature's toolbox for white biotechnology. Appl Microbiol Biotechnol 80:381-388

Cho WK, Chen XY, Chu H, Rim Y, Kim S, Kim ST, Kim, SW, Park ZY, Kim JY (2009) Proteomic analysis of the secretome of rice calli. Physiol Plant 135:331–341

Emanuelsson O, Brunak S, von Heijne G, Nielsen H (2007) Locating proteins in the cell using TargetP, SignalP and related tools. Nat Protoc 2:953-971

Greenbaum D, Luscombe NM, Jansen R, Qian J, Gerstein M (2001) Interrelating different types of genomic data, from proteome to secretome: 'oming in on function. Genome Res 11:1463–1468

Hathout Y (2007) Approaches to the study of the cell secretome. Expert Rev Proteomics 4:239-248

Isaacson T, Rose JKC (2006) The plant cell wall proteome, or secretome. In: Finnie C (ed) Plant Proteomics. Annual Plant Reviews Series. Blackwell Publishing 28:185-209

Jones RL, Robinson DG (1989) Protein Secretion in Plants. Tansley Review No. 17. New Phytologist 111:567-597

Jung YH, Jeong SH, Kim SH, Singh R, Lee JE, Cho YS, Agrawal GK, Rakwal R, Jwa NS (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

Käll L, Krogh A, Sonnhammer EL (2004) A combined transmembrane topology and signal peptide prediction method. J Mol Biol 338:1027-1036

Kamoun S (2009) The Secretome of plant-associated fungi and oomycetes. In: Deising VH (ed) Plant Relationships, 2nd Edition, The Mycota. Springer-Verlag, Berlin Heidelberg. pp 173-180

Kusumawati L, Imin N, Djordjevic MA (2008) Characterization of the secretome of suspension cultures of Medicago species reveals proteins important for defense and development. J Proteome Res 7:4508–4520

Lopez-Casado G, Urbanowicz BR, Damasceno CMB, Rose JKC (2008) Plant glycosyl hydrolases and biofuels: a natural marriage. Current Opinion Plant Biol 11:329-337

Marchler-Bauer A, Anderson JB, Chitsaz F Derbyshire MK, DeWeese-Scott C, Fong JH, Geer LY, Geer RC, Gonzales NR, Gwadz M, He S, Hurwitz DI, Jackson JD, Ke Z, Lanczycki CJ, Liebert CA, Liu C, Lu F, Lu S, Marchler GH, Mullokandov M, Song JS, Tasneem A, Thanki N, Yamashita RA, Zhang D, Zhang N, Bryant SH (2009) CDD: specific functional annotation with the Conserved Domain Database. Nucleic Acids Res. 37:D205-210

McCarthy FM, Wang N, Magee GB, Williams WP, Luthe DS, Burgess SC (2006) AgBase: a functional genomics resource for agriculture. BMC Genomics. 7:229

Min XJ (2010) Development of computational protocols for secreted protein prediction in different eukaryotes. J Proteomics Bioinformatics 4:143-147

Ranki H, Sopanen T (1984) Secretion of alpha-amylase by the aleurone layer and the scutellum of germinating barley grain. Plant Physiol 75:710-715

Sampedro J, Cosgrove DJ (2005) The expansin superfamily. Genome Biol 6:242

Simpson JC, Mateos A, Pepperkok R (2007) Maturation of the mammalian secretome. Genome Biol 8:211

Sottomayor M, Barceló AR (2004) Plant peroxidases and phytochemistry – foreword. Phytochemistry Reviews 3: 1–2

The UniProt Consortium (2010) The universal protein resource (UniProt) in 2010. Nucleic Acids Res 38:D142-148

Tjalsma H, Bolhuis A, Jongbloed JD, Bron S, van Dijl JM (2000) Signal peptide-dependent protein transport in *Bacillus subtilis*: a genome-based survey of the secretome. Microbiol Mol Biol Rev 64:515-547

Tran HT, Plaxton WC (2008) Proteomic analysis of alterations in the secretome of *Arabidopsis thaliana* suspension cells subjected to nutritional phosphate deficiency. Proteomics 8:4317–4326