

Genome-Wide Identification and Analysis of Genes Encoding Proteolytic Enzymes in Pineapple

Ching Man Wai¹ · Brian Powell² · Ray Ming¹ · Xiang Jia Min³

Received: 26 December 2015 / Accepted: 10 May 2016
© Springer Science+Business Media New York 2016

Abstract Pineapple, *Ananas comosus*, is an economically important fruit crop. Recently its genome was completely sequenced and a total of 27,024 protein coding genes were predicted. Using a set of well evaluated bioinformatics tools we have predicted the protein subcellular locations and comparatively analyzed the protein conserved domains of the predicted proteomes in pineapple, *Oryza sativa* (rice), *Sorghum bicolor* (sorghum), and *Brachypodium distachyon*. Our analysis revealed that ~24–26 % of proteins were located in nucleus, 17–21 % in cytosol, 9–11 % in chloroplast, and 8–11 % proteins were secreted in these monocot plants. The secretomes in the four species were analyzed comparatively and a large number of secreted glycosyl hydrolases were identified. As pineapple proteolytic enzymes, known as bromelains, have been used for medical treatments, we focused on genome-wide identification and analysis of pineapple genes encoding proteases. A total of 512 pineapple genes encoding putative proteolytic enzymes were identified, with 152 secreted, 74 localized in cytosol, 67 in nucleus, 60 in

chloroplast, 18 in mitochondria, and the remaining in other subcellular locations. The top large protease families in pineapple were papain family cysteine protease (62 genes), peptidase S8 family (56 genes), aspartyl protease family (38 genes), and serine carboxypeptidase (33 genes). Gene expression analysis revealed that among 512 protease genes 432 were expressed in various tissues and 72 genes were differentially expressed. The highly expressed protease genes were identified including 7 papain family cysteine proteases. The protease genes with the predicted protein subcellular locations will facilitate the efforts for examining their biological roles in pineapple growth and development and for expressing the recombinant proteases for medical use. The information of protein subcellular location of all plant species can be accessed at the PlantSecKB website (<http://proteomics.yzu.edu/secretomes/plant.php>).

Keywords Bromelain · Conserved domain · Gene expression · Proteolytic enzyme · Protease · Secreted protein · Secretome · Subcellular locations

Communicated by: Paulo Arruda

Electronic supplementary material The online version of this article (doi:10.1007/s12042-016-9172-5) contains supplementary material, which is available to authorized users.

✉ Xiang Jia Min
xmin@ysu.edu

¹ Department of Plant Biology, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

² Department of Computer Science and Information Systems, Youngstown State University, Youngstown, OH 44555, USA

³ Department of Biological Sciences, Center for Applied Chemical Biology, Youngstown State University, Youngstown, OH 44555, USA

Abbreviations

BLAST	basic local alignment search tool.
ER	endoplasmic reticulum.
GO	Gene ontology.
FPKM	Fragments Per Kilobase exon per Million reads mapped.
rpsBLAST	reversed position specific BLAST.

Introduction

Pineapple, *Ananas comosus* L., is an important tropical and subtropical fruit plant. It possesses a crassulacean acid

metabolism (CAM) photosynthesis pathway, i. e., carbon dioxide is fixed as malate during the night and the carbon is re-fixed using C3-pathway during the day (Bartholomew and Kadzimin 1977; Bartholomew and Malézieux 1994). Its fruit ripening process is known to be non-climacteric, i. e., ripening without ethylene and respiration bursts. Its fruits and juice can be consumed directly. Its products also include bromelain which was used as folk medicine and a culinary ingredient as a meat tenderizer (Taussig and Batkin 1988). Bromelain belongs to a group of protein digesting enzymes obtained commercially from the fruit or stem of pineapple (Pavan et al. 2012 for review). Stem bromelain is a mixture of different proteolytic enzymes and also contains other components, in small quantity, like phosphatase, glucosidase, peroxidase, cellulase, escharase, and several protease inhibitors (Heinicke and Gortner 1957). In vitro and in vivo studies demonstrate that bromelain exhibits various fibrinolytic, antiedematous, antithrombotic, and anti-inflammatory activities (Pavan et al. 2012; Meiser et al. 2014).

Stem bromelain (EC 3.4.22.32) was isolated from the stem of pineapples and was characterized as a complex of cysteine proteases (Taussig and Batkin 1988; Arshad et al. 2014). Cysteine proteases are involved in various physiological and developmental processes in plants including programmed cell death during organ senescence and tracheary element differentiation (Van der Hoorn 2008 for review). Pineapple stem bromelain, a cysteine protease, is an inhibitor of fungal plant pathogens (López-García et al. 2012). Overexpression of the pineapple fruit bromelain gene (Accession number: BAA21849) in transgenic Chinese cabbage (*Brassica rapa*) results in enhanced resistance to bacterial soft rot (Jung et al. 2008). Another cysteine proteinase (AcCP2) gene from pineapple fruit was found to be gradually increased in its expression during fruit development until maturity, and the overexpression of the gene in *Arabidopsis thaliana* improved the resistance to fungal pathogen of *Botrytis cinerea* (Wang et al. 2014). The interaction of fruit bromelain (EC 3.4.22.33) and cystatins, inhibitors of cysteine proteinases, regulates the fruit development and ripening, and directly correlates to the resistance to blackheart development in pineapple fruits during postharvest chilling stress (Neuteboom et al. 2009; Raimbault et al. 2013a). A minor form of pineapple stem cysteine protease, named as ananain (EC 3.4.22.31), was isolated and found that it differs from stem and fruit bromelains in being inhibited by chicken cystatin (Lee et al. 1997; Carter et al. 2000). Recently a novel aspartic acid protease gene (AcAP1) from pineapple fruit was characterized and found the gene was involved in postharvest chilling stress resistance in pineapple fruits (Raimbault et al. 2013b). One common feature among all these characterized pineapple enzymes

including bromelains, cystatins, and aspartic acid protease, we noticed, is that they are all classic secreted enzymes, i. e., possessing a secretory signal peptide.

Recently the complete pineapple genome was sequenced and annotated (Ming et al. 2015). A total of 27,024 protein coding genes from the genome were predicted. The protein dataset provides an unprecedented opportunity for proteome-wide identification of all proteolytic enzymes from pineapple. Identifying these genes and proteins will facilitate further characterization of these genes and their products in the processes of pineapple growth, development, and fruit storage, as well as expression and characterization of recombinant proteins for commercial applications. The effort for expressing recombinant protein has been made with pineapple stem bromelain (Amid et al. 2011). Recently the pineapple stem bromelain was successfully expressed in a prokaryotic system and proved that the recombinant bromelain protein had the bactericidal property (George et al. 2014).

A plant synthesizes tens of thousands of different proteins, and the synthesized proteins are transported to different subcellular locations to play their biological roles. Some of these proteins are secreted outside of the cell, including cell walls and the extracellular space. These secreted proteins of a proteome are collectively called a secretome (Agrawal et al. 2010; Lum and Min 2011). Plant secretome consists of primarily cell wall proteins, proteins involved in carbon and cell wall metabolism, and extracellular enzymes and signal molecules involved in defense of pathogens (Isaacson and Rose 2006). We analyzed all manually curated and annotated secreted plant proteins in the UniProt Swiss-Prot database and found that the majority of identified secreted proteins (>93 %) are classical secreted proteins that contain a secretory signal peptide which can be detected by a prediction tool (Lum and Min 2011). Recently we have constructed the PlantSecKB, the plant secretome and subcellular proteome knowledgebase (<http://proteomics.yzu.edu/secretomes/plant/index.php>), using all available plant protein sequences (over 1.4 million entries) in UniProtKB (Lum et al. 2014) and the protein sequences predicted from the newly sequenced genome *Nelumbo nucifera* (Gaertn.) (Lum et al. 2013). In this study, we present the protein subcellular location prediction and protein conserved domain analysis in pineapple with an attempt to identify all genes encoding proteolytic enzymes. Particularly, the secreted proteolytic enzymes are potential targets for commercial applications. This work is expected to provide a solid molecular resource for future elucidating the roles of proteases in pineapple growth and development, and also for exploring the potential of expression and characterization of recombinant proteins for commercial applications in medicine and the food industry (González-Rábade et al. 2011).

Results

Overview of Proteomes in Different Monocot Species

To provide an overview of proteomes in pineapple and other three monocot species we mapped the proteins to Gene Ontology (GO) classification. Though there were some variations in these categories of GO classification, the overall distributions of the GO categories in these four monocot plants did not show much obvious difference

(Supplementary Table 1). Since the GO categories contained complex hierarchy structure and some proteins did not have GO classification and some had more levels of GO categories, the subtle differences in the proteomes in these species were not detected. Thus we further compared the conserved domains and protein families of the proteomes using rpsBLAST against the Conserved Domain dataset (see method section). The protein families having 50 or more members were listed in Table 1. The complete list was provided as Supplementary Table 2.

Table 1 Conserved domain (CD) and protein family distributions in the proteomes of four monocots

	Pineapple	Brachypodium	Rice	Sorghum	
cd14066	419	667	803	625	STKc_IRAK; Catalytic domain of the Serine/Threonine
PLN03077	262	197	196	187	PLN03077; Protein ECB2; Provisional.
PLN00113	192	228	337	270	PLN00113; leucine-rich repeat receptor-like protein
pfam00931	162	224	441	289	NB-ARC; NB-ARC domain.
cd00200	158	194	156	150	WD40; WD40 domain
pfam13639	145	196	153	192	zf-RING_2; Ring finger domain.
sd00004	144	146	155	165	PPR; Pentatricopeptide repeat
smart00380	99	123	128	129	AP2; DNA-binding domain in plant proteins
pfam10536	85	17	20	86	PMD; Plant mobile domain.
cd00693	82	142	145	148	secretory_peroxidase; Horseradish peroxidase and related
PLN03091	75	82	77	86	PLN03091; hypothetical protein; Provisional.
cd01837	72	118	128	115	SGNH_plant_lipase_like; SGNH_plant_lipase_like
pfam02365	72	109	135	118	NAM; No apical meristem (NAM) protein.
PLN03081	71	53	58	53	PLN03081; pentatricopeptide (PPR) repeat-containing
PLN03218	71	76	71	80	PLN03218; maturation of RBCL 1; Provisional.
pfam02458	63	80	107	74	Transferase; Transferase family.
cd04852	56	62	64	58	Peptidases_S8_3; Peptidase S8 family domain
TIGR01557	56	75	64	71	unnamed_protein_product; myb-like DNA-binding domain
PHA03247	55	31	44	112	PHA03247; large tegument protein UL36; Provisional.
cd00143	52	79	53	62	PP2Cc; Serine/threonine phosphatases
cd13999	52	67	55	57	STKc_MAP3K-like; Catalytic domain of Mitogen-Activated
pfam03140	50	49	73	41	DUF247; Plant protein of unknown function.
Total count	18268	23782	22470	23796	
CD type	5063	5079	4903	5003	
Total proteins	27024	29942	34077	32796	

The complete list of the conserved domain and protein family distribution can be found in the Supplementary Table 2

Overall, as the number of proteins encoded by the pineapple genome is fewer than the numbers of proteins in other three monocots, the numbers of proteins in many conserved protein domain families are slightly fewer than in other species. This might be due to fact that the pineapple genome has one fewer ancient whole-genome duplication event than the grass genomes (Ming et al. 2015). Some domains in pineapple proteome showed strikingly lower number of members such as the STKc_IRAK domain (Catalytic domain of the Serine/Threonine kinases, Interleukin-1 Receptor Associated Kinases and related STKs), secretory peroxidase family, and the NAM domain (No apical meristem protein). In the STKc_IRAK domain family, pineapple proteome has 419 members and there are 667 members in Brachypodium, 803 members in rice, and 625 members in sorghum. In the secretory peroxidase family, there are only 82 in pineapple proteome, while there are more than 140 members in other three species. In the NAM domain, pineapple proteome has 72 members while there are 109 members in Brachypodium, 135 members in rice, and 118 members in sorghum (Table 1). Interestingly, we also noted that some conserved domain families had more members in pineapple proteome than proteomes in other three species, such as the Provisional Protein ECB2 (a pentatricopeptide repeat protein, PLN03077) and another provisional pentatricopeptide (PPR) repeat-containing protein domain (PLN03081). The comparative analysis also revealed that rice proteome had 441 proteins with NB-ARC domain, much higher than other species had. The NB-ARC domain is a novel signaling motif shared by plant resistance gene products and regulators of cell death in animals (van der Biezen et al. 1998). Collectively, these genes and their functional

products make each species common in some aspects and unique in others in their phenotypes.

Protein Subcellular Locations: Secretomes and other Subcellular Proteomes

Using the protocol we developed for prediction of plant subcellular locations (Lum et al. 2014; Min 2010), the subcellular locations for all pineapple proteins and three monocot plants including rice, sorghum and Brachypodium were predicted. The data from all tools for proteins in each species are integrated into one file per species and can be downloaded using the URL provided in the method section. The detailed output from each prediction tool can be obtained by searching the PlantSecKB using a protein ID (Lum et al. 2014). The distributions of each category of subcellular proteomes in each species were summarized in Table 2. Nucleus, cytosol and chloroplast represented three larger compartments for protein localization (Table 2). Secreted proteins accounted for 8.1–10.6 % of the whole proteome set in the species analyzed, which were slightly higher than we reported previously (Lum et al. 2014), as in this work, the secreted proteins included both the predicted highly likely secreted proteins and likely secreted proteins. A total of 2, 183 secreted proteins were identified in the pineapple proteome. Noticeably, pineapple had the smallest number of protein coding genes in these species compared, also had the smallest number of genes encoding secreted proteins among these species (Table 2). However, the number of predicted proteins located in the chloroplast in pineapple was close to the numbers of chloroplast proteins in other three monocots. It also needs to be noted that transmembrane proteins are proteins with one or more predicted

Table 2 Subcellular proteome distribution in pineapple and three other monocot plants

	Pineapple (%)	Brachypodium (%)	Rice (%)	Sorghum (%)
Chloroplast	2931 (10.8)	2993 (10.0)	3030 (8.9)	2997 (9.1)
Cytoskeleton	158 (0.6)	228 (0.8)	213 (0.6)	182 (0.6)
Cytosol	4682 (17.3)	5392 (18.0)	6620 (19.4)	6730 (20.5)
ER	186 (0.7)	254 (0.8)	211 (0.6)	208 (0.6)
Golgi	66 (0.2)	53 (0.2)	43 (0.1)	39 (0.1)
Mitochondria	1272 (4.7)	1338 (4.5)	1846 (5.4)	1627 (5.0)
Nucleus	7061 (26.1)	7508 (25.1)	8323 (24.4)	8123 (24.8)
Peroxisome	62 (0.2)	74 (0.2)	64 (0.2)	111 (0.3)
Plasma membrane	1384 (5.1)	1662 (5.6)	1541 (4.5)	1518 (4.6)
Secreted	2183 (8.1)	3175 (10.6)	3446 (10.1)	3185 (9.7)
Transmembrane	1294 (4.8)	1583 (5.3)	1728 (5.1)	1564 (4.8)
Vacuole	215 (0.8)	269 (0.9)	317 (0.9)	300 (0.9)
Two locations	165 (0.6)	137 (0.5)	124 (0.4)	202 (0.6)
Unassigned	5365 (19.9)	5276 (17.6)	6571 (19.3)	6010 (18.3)
Total	27024	29942	34077	32796

transmembrane domains but not being assigned to a subcellular location (Table 2). In other subcellular locations except cytosol and secreted proteins, there were membrane proteins associated with each subcellular compartment, and these membrane proteins and non-membrane proteins were able to be searched and downloaded in the PlantSecKB (Lum et al. 2014). The predicted subcellular location features of these proteins in pineapple and other species are particularly useful for designing experiments for further characterizing these proteins.

Secretome Analysis Using Gene Ontology (GO) and Conserved Domain Classification

We further analyzed the secretome in each species using GO molecular function (Table 3). Plant secreted proteins play important roles in many biological processes including metabolic and catabolic process, cellular component organization and organismal development, signal transduction and response to stress, etc. (Agrawal et al. 2010). The main types of molecular functions are binding (~40%), hydrolase activity (17–19%) and transferase activity (9–10%), and other catalytic activity (12%) (Table 3). The overall distribution of each category of molecular function and biological process of secreted proteins in each species was similar in all monocot species studied in the work, though there were fewer counts

in each category in pineapple due to its smaller secretome size (Table 3).

Protein family analysis revealed that the members of each protein family within secretomes among plant species were different (Supplementary Table 3). The top 6 secreted protein families with more than 30 members in pineapple were secretory peroxidase (66), leucine-rich repeat receptor-like protein kinase (57); SGNH plant lipase like family (54), Cupin 1 (42), NB-ARC domain (33), Catalytic domain of the Serine/Threonine (31) (Table 4). As pineapple has a relatively smaller proteome than the grass species, likely due to its one fewer whole genome duplication than the grass species (Ming et al. 2015), the members of secreted proteins in most protein families were lower than the numbers in the grass species examined. However, noticeably, the smaller secretome size in pineapple does not mean it has lower numbers within all protein families. In fact, it has more members than other species in the some of the protein families, for example, pineapple has 29 members in secreted aspartyl protease, while the grass species have only 6 or 7 members (Table 4). Pineapple has 5 members of galactosyltransferase (Pfam01762) while the grass species has only 1 or 2 members. In addition, each species has their own species specific gene families in the secretomes, for example, there were 197 secreted proteins belonging to 186 protein families unique in pineapple proteome (Supplementary Table 3). We also noticed that

Table 3 Classification based on Gene Ontology (GO) molecular function in secretomes of different species

	Pineapple		Brachypodium		Rice		Sorghum		
	count	%	count	%	count	%	count	%	
GO:0005488	468	23.7	644	24.3	636	23.4	704	25.4	binding
GO:0016787	371	18.8	506	19.1	463	17.0	497	17.9	hydrolase activity
GO:0003824	254	12.9	333	12.6	318	11.7	329	11.9	catalytic activity
GO:0016740	198	10.0	239	9.0	252	9.3	255	9.2	transferase activity
GO:0000166	140	7.1	168	6.3	206	7.6	215	7.8	nucleotide binding
GO:0016301	86	4.4	99	3.7	137	5.0	126	4.6	kinase activity
GO:0003674	77	3.9	111	4.2	90	3.3	112	4.0	Molecular function
GO:0004871	60	3.0	83	3.1	119	4.4	90	3.3	signal transducer activity
GO:0004872	55	2.8	77	2.9	109	4.0	86	3.1	receptor activity
GO:0005515	50	2.5	75	2.8	99	3.6	76	2.7	protein binding
GO:0030246	42	2.1	67	2.5	70	2.6	72	2.6	carbohydrate binding
GO:0003677	29	1.5	52	2.0	37	1.4	47	1.7	DNA binding
GO:0008289	28	1.4	31	1.2	29	1.1	29	1.0	lipid binding
GO:0003723	22	1.1	34	1.3	27	1.0	14	0.5	RNA binding
GO:0030234	20	1.0	36	1.4	41	1.5	40	1.4	enzyme regulator activity
Others	74	3.7	95	3.6	89	3.3	77	2.8	
Total	1974		2650		2722		2769		

Table 4 Protein families in secretomes of different species

	Pineapple	Brachypodium	Rice	Sorghum	
cd00693	66	136	114	131	cd00693, secretory_peroxidase, Horseradish peroxidase and related
PLN00113	57	71	86	69	PLN00113, PLN00113, leucine-rich repeat receptor-like protein kinase;
cd01837	54	101	97	81	cd01837, SGNH_plant_lipase_like, SGNH_plant_lipase_like, a plant
pfam00190	42	37	8	43	pfam00190, Cupin_1, Cupin. This family represents the conserved
pfam00931	33	18	53	37	pfam00931, NB-ARC, NB-ARC domain.
cd14066	31	63	84	58	cd14066, STKc_IRAK, Catalytic domain of the Serine/Threonine
PLN03146	29	6	6	7	PLN03146, PLN03146, aspartyl protease family protein; Provisional.
cd04852	28	57	39	49	cd04852, Peptidases_S8_3, Peptidase S8 family domain,
pfam00450	26	62	42	48	pfam00450, Peptidase_S10, Serine carboxypeptidase.
pfam00112*	26	34	27	36	pfam00112, Peptidase_C1, Papain family cysteine protease.
pfam00332	22	43	40	33	pfam00332, Glyco_hydro_17, Glycosyl hydrolases family 17.
cd05476	22	23	32	28	cd05476, pepsin_A_like_plant, Chloroplast Nucleoids DNA-binding
cd02176	21	30	25	31	cd02176, GH16_XET, Xyloglucan endotransglycosylase, member of
pfam03018	16	32	35	30	pfam03018, Dirigent, Dirigent-like protein. This family contains a
TIGR03389	15	30	13	22	TIGR03389, laccase_11, laccase, plant. Members of this protein
pfam01190	16	29	41	24	pfam01190, Pollen_Ole_e_I, Pollen proteins Ole e I like.
PLN02687	5	25	39	37	PLN02687, PLN02687, flavonoid 3'-monooxygenase.
PLN03023	9	25	13	32	PLN03023, PLN03023, Expansin-like B1; Provisional.
cd04216	10	22	18	24	cd04216, Phycocyanin, Phycocyanins are plant blue or type I copper
cd00010	12	22	18	15	cd00010, AAI_LTSS, AAI_LTSS: Alpha-Amylase Inhibitors (AAI), Lipid
cd01958	11	21	27	22	cd01958, HPS_like, HPS_like: Hydrophobic Protein from Soybean
PLN00193	8	21	18	28	PLN00193, PLN00193, expansin-A; Provisional.
pfam13947	5	9	23	14	pfam13947, GUB_WAK_bind, Wall-associated receptor kinase
cd02877	5	13	22	19	cd02877, GH18_hevamine_XipI_class_III, This conserved domain family
pfam01657	9	10	21	13	pfam01657, Stress-antifungal, Salt stress response/antifungal. This
cd05472	13	19	15	37	cd05472, cnd41_like, Chloroplast Nucleoids DNA-binding Protease,
pfam04398	9	19	15	21	pfam04398, DUF538, Protein of unknown function, DUF538. This
pfam00232	11	17	4	21	pfam00232, Glyco_hydro_1, Glycosyl hydrolase family 1.

Table 4 (continued)

	Pineapple	Brachypodium	Rice	Sorghum	
cd15795	6	17	11	20	cd15795, PMEI-Pla_a_1_like, Pollen allergen Pla a 1 and similar

The complete list of the protein family distribution in secretomes can be found in the Supplementary Table 5

* The entries of pfam00112 include pfam00112, cd02248, cd02620, smart00645, and cd02619, as they all are classified as Peptidase_C1 family

there are a large number of secreted glycosyl hydrolases classified into 7–9 protein families in these plant species, with 49 in pineapple, 55 in rice, 66 in sorghum, and 82 in Brachypodium (Supplementary Table 3). These diverse secreted plant glycosyl hydrolases may have valuable applications in modifying plant cell wall for bio-energy production (Lepoz-Casado et al. 2008). All together these secreted proteins play important roles in cell wall formation, biotic and abiotic stress response and defense, and metabolism during plant growth and development (Tran and Plaxton 2008; Agrawal et al. 2010).

Proteolytic Enzymes

As our focus was on the proteolytic enzymes, we retrieved the protein domain families having the terms of peptidase, protease or proteinase (Barrett and McDonald 1986) but excluded entries having a term of “inhibitor”. The results for protease families having 10 members or more in at least one species were summarized in Table 5. The complete list was provided in Supplementary Table 4. In the proteome of each species there are a total of more than 500 proteolytic enzymes, with 512 in pineapple, 546 in rice, 535 in sorghum and 646 in Brachypodium (Supplementary Table 5 for pineapple, others can be downloaded). The larger number of proteases in Brachypodium was worth noting in considering its relative smaller proteome in comparing the proteome sizes of rice and sorghum. Pineapple proteome has 62, slightly more number of Peptidase_C1 (papain family cysteine protease) than other species, with 58 in rice and 45 in both sorghum and Brachypodium. Pineapple bromelain was characterized as a complex of cysteine proteases (Taussig and Batkin 1988; de Lencastre Novaes et al. 2016 for review). However, pineapple has a significantly higher number of aspartyl protease family proteins (PLN03146), 38 members in pineapple, while other species has less than 10 members. One pineapple aspartic acid protease gene (AcAP1) was recently examined and found that the gene was involved in postharvest chilling stress resistance in pineapple fruits (Raimbault et al. 2013b). More interestingly, pineapple proteome has 15 members of RVP_2 (single domain retroviral aspartyl protease), while there is no protein having this domain in other species. Thus the origin of RVP_2 proteins and their functional

significance in pineapple is worthy further examined. Pineapple has only 1 member of Trypsin_2 (Trypsin-like peptidase domain), while there are 17–22 members in other species. We also identified the protease inhibitors in difference species with a total of 10 members in pineapple, 38 members in rice, 25 members in sorghum and 41 members in Brachypodium (Supplementary Table 2). Plant protease inhibitors are involved in plant defense of pathogens and insects (Habib and Fazili 2007). Certainly the biological significances of proteases and protease inhibitors as well as their interactions warrant further detailed examination in these plants.

Within the protease families we further identified the secreted proteases as so far analyzed proteases in literature were all secreted and played important biological roles (see introduction). The total numbers of secreted proteases were 152 in pineapple, 161 in rice, 218 in sorghum, and 222 in Brachypodium (Table 5). The large secreted protease families are papain family cysteine protease (Peptidase_C1), peptidase S8 family (Peptidases_S8_3), provisional aspartyl protease family (PLN03146), serine carboxypeptidase (Peptidase_S10), and chloroplast nucleoids DNA-binding protease (cnd41_like) (Table 5). The most noticeable difference is the number of secreted aspartyl protease family, there are 29 members in pineapple and only 6–7 members in other species. Pineapple proteases were also localized in other subcellular compartments including 74 localized in cytosol, 60 in chloroplast, 67 in nucleus, and 18 in mitochondria, and a small number of proteases in other various locations.

RNA-seq Expression Analysis of Genes Involved in Fruit Ripening and Proteases

We presented some detailed analysis of gene expression in another report (Wai et al. 2016), our current analysis focused on expression of genes encoding proteases. There were a total of 12,768 genes expressed having ≥ 10 FPKM (Fragments Per Kilobase of exon model per Million mapped reads) in at least one tissue. Based on at least two fold changes of the \log_2 values of FPKM, pairwise tissue gene expression comparison revealed that 3,502 of them were differentially expressed in different tissues, and

Table 5 Proteases include secreted proteases identified from the proteomes in each species

CD ID	Pineapple	Rice	Sorghum	Brachypodium	Conserved domain description
	All (Sec)	All (Sec)	All (Sec)	All (Sec)	
pfam00112*	62 (26)	58 (27)	45 (36)	45 (34)	Peptidase_C1; Papain family cysteine protease
cd04852	56 (28)	64 (39)	58 (49)	62 (57)	Peptidases_S8_3; Peptidase S8 family domain
PLN03146	38 (29)	8 (6)	9 (7)	7 (6)	PLN03146; aspartyl protease family protein; Provisional
pfam00450	33 (26)	52 (42)	54 (48)	75 (62)	Peptidase_S10; Serine carboxypeptidase
cd05472	21 (13)	25 (15)	44 (37)	26 (19)	cnd41_like; Chloroplast Nucleoids DNA-binding Protease
pfam08284	15 (1)	0 (0)	0 (0)	0 (0)	RVP_2; Retroviral aspartyl protease. Single domain
TIGR01241	13 (0)	4 (0)	6 (0)	12 (0)	ATP-dependent_zinc_metalloprotease_FtsH; ATP-dependent
cd07017	12 (1)	5 (0)	5 (0)	8 (0)	S14_ClpP_2; Caseinolytic protease (ClpP)
pfam02338	10 (0)	13 (0)	10 (0)	20 (0)	OUT; OTU-like cysteine protease
COG0612	10 (0)	8 (0)	6 (0)	11 (0)	PqqL; Predicted Zn-dependent peptidases
pfam13365	1 (0)	22 (0)	18 (0)	17 (1)	Trypsin_2; Trypsin-like peptidase domain
pfam02902	4 (0)	12 (1)	20 (1)	6 (0)	Peptidase_C48; Ulp1 protease family
TIGR02037	8 (0)	11 (1)	7 (1)	19 (1)	Probable_periplasmic_serine_protease_do/HhoA-like
cd02667	1 (0)	11 (0)	5 (0)	4 (0)	Peptidase_C19K; A subfamily of Peptidase C19
pfam05903	7 (0)	10 (0)	10 (0)	14 (0)	Peptidase_C97; PPPDE putative peptidase domain
pfam00656	5 (0)	9 (0)	12 (0)	12 (0)	Peptidase_C14; Caspase domain
COG0542	7 (0)	5 (0)	7 (0)	15 (3)	clpA; ATP-binding subunits of Clp protease and DnaK/DnaJ
cd02661	9 (1)	8 (1)	7 (1)	13 (2)	Peptidase_C19E; A subfamily of Peptidase C19
TIGR02227	5 (1)	8 (0)	9 (0)	11 (0)	Inactive_signal_peptidase_IA; signal peptidase I
COG1506	5 (2)	5 (0)	4 (0)	10 (3)	DAP2; Dipeptidyl aminopeptidases/acylaminoacyl-peptidases
Total	512 (152)	546 (161)	535 (218)	646 (222)	

The complete list of protease families is shown as Supplementary Table 3. Sec: secreted

* The entries of pfam00112 include pfam00112, cd02248, cd02620, smart00645, and cd02619, as they all are classified as Peptidase_C1 family

532 of them were predicted to encode secreted proteins. Comparing gene expression in young fruits with ripening fruits revealed that in the ripening fruits 931 genes were up-regulated including 167 of them encoding secreted proteins, and 189 were down regulated with 20 of them encoding secreted proteins. These up-regulated genes

encoded diverse families of enzymes including lipase, glycosyl hydrolases, proteases, peroxidases, and others, which were involved in fruit ripening process (Supplementary Table 6).

We identified 512 genes encoding putative proteases in pineapple. Among them 159 were lowly expressed with < 10

FPKM in all tissues, 273 were expressed (having ≥ 10 FPKM in at least one tissue), the remaining genes were not expressed. Among the 273 expressed genes encoding proteases, there are 71 secreted proteins, 42 proteins localized in cytosol, 42 proteins in chloroplast, 32 proteins in nucleus, 10 proteins in mitochondria, 7 proteins in ER, and the remaining localized in other subcellular compartments or the locations not predicted (Supplementary Table 7). In addition, we also found that 72 protease coding genes were differentially expressed in different tissues. We also found that among the proteins encoded by the 273 expressed protease genes there were 25 papain family cysteine proteases (pfam00112, Peptidase_C1), 18 serine carboxypeptidase (pfam00450, Peptidase_S10), 11 having peptidase S8 family domain (cd04852, Peptidases_S8_3), 10 chloroplast nucleoids DNA-binding proteases (cd05472, cnd41_like), 9 OTU-like cysteine proteases (pfam02338, OUT), etc. (Supplementary Table 7).

The genes having > 100 FPKM were listed in Table 6. The top 9 highly expressed genes, having > 400 FPKM, included 7 genes encoding papain family cysteine protease, i. e., bromelain, with 6 of them encoding secreted cysteine proteases (Table 6). Even among these 7 bromelain genes, they had different expression levels and some showed differential expression among different tissues. For example, Aco004356 gene encoding a cysteine protease was extremely highly expressed in both young fruit and ripening fruit with > 7000 FPKM, while also highly expressed in flower and leaf tissues, relatively modestly expressed in root tissue. Aco017317 gene was expressed lowly in young fruit and highly expressed in leaf, flower and ripening fruit tissues. Aco004355 gene was highly expressed in all tissues except root tissues (only 2 FPKM), while Aco013510 was highly expressed in all tissues sampled (Table 6).

Discussion

Pineapple is an economically important fruit crop and possesses CAM photosynthetic pathway. Its complete genome sequencing provides a solid resource for identifying genes involved in CAM pathway and engineering drought tolerant crop species (Ming et al. 2015). As pineapple bromelains can be used for medical treatments (Meiser et al. 2014), also proteolytic enzymes were involved in fruit ripening (Neuteboom et al. 2009; Raimbault et al. 2013a, b), our focus of this work was to identify all proteolytic enzymes from predicted pineapple proteome. Comparative analysis with other monocot species including rice, sorghum, and *Brachypodium* revealed that, though pineapple had a relatively smaller proteome, several protein families including the provisional protein ECB2 family (PLN03077), pentatricopeptide (PPR) repeat-containing protein domain family (PLN03081), and aspartyl

protease family (PLN03146) were expanded (Table 1 and 4). The information on subcellular locations of all plant proteins downloaded from UniProtKB and predicted proteins from newly sequenced genomes of sacred locus and pineapple was available in PlantSecKB and can be used for further comparative analysis (Lum et al. 2013; Lum 2014).

Plant proteases are key regulators in various biological processes, including meiosis, gametophyte survival, embryogenesis, seed coat formation, cuticle deposition, epidermal cell fate, stomata development, chloroplast biogenesis, and local and systemic defense responses (Van der Hoorn 2008 for review). We identified 512 genes encoding putative proteolytic enzymes with 150 of them were secreted (Table 5), among them 432 were expressed with 273 genes having ≥ 10 FPKM in at least one tissue. These highly expressed genes included papain family cysteine proteases (6 secreted out of 7), M20 peptidase, serine carboxypeptidase, etc. (Table 6). Plastids, mitochondria, and peroxisomes are key organelles in plants, and the plant organellar proteases play important roles in degrading unwanted proteins within organelles (van Wijk 2015). The predicted organellar proteases (Supplementary Table 7) in our analysis will be useful for further elucidating their biological roles. In our analysis we also identified 931 up-regulated genes in the ripening fruits relative to the young fruits (Supplementary Table 7). In this work, we did not measure gene expression in stem tissues. In future, gene expression measurements coupling with proteomic analysis in both stem and fruit tissues will be needed for identifying the bioactive bromelain protein sequences for recombinant protein production. Thus the work along with the information in PlantSecKB is expected to provide a resource that will help the community to further design experiments to characterize the genes and proteins to understand the biology of pineapple and explore the potential for commercial applications.

Methods

Data Collection

A total of 27,024 protein sequences were predicted from the genome sequences of pineapple (Ming et al. 2015), the data were deposited into GenBank and can be downloaded from our website (URL). For comparative analysis complete proteome data in other plant species including *Oryza sativa* (subsp. japonica) (rice, UP0000077520), *Sorghum bicolor* (sorghum, UP00000768), and *Brachypodium distachyon* (purple false brom, UP000008810) were downloaded from the UniProt database (The UniProt Consortium 2015) (<http://www.uniprot.org/taxonomy/complete-proteomes>).

Table 6 Gene expression levels of proteases in different tissues of pineapple

ID	Leaf	Flower	Root	Fruit1_5	Fruit6_8	Location	Conserved domain
Aco004356.1	4917	4756	292	9749	7462	S	pfam00112, Peptidase_C1, Papain family cysteine protease.
Aco017317.1	2748	2746	190	18	1840	E	pfam00112, Peptidase_C1, Papain family cysteine protease.
Aco013510.1	1777	1676	1236	1444	1968	S	pfam00112, Peptidase_C1, Papain family cysteine protease.
Aco004355.1	1046	1164	2	2301	1616	-	pfam00112, Peptidase_C1, Papain family cysteine protease.
Aco028299.1	649	747	0	119	551	N	cd08012, M20_ArgE-related, M20 Peptidases with similarity to
Aco011776.1	602	612	167	121	443	S	pfam00450, Peptidase_S10, Serine carboxypeptidase.
Aco011478.1	442	403	529	467	419	S	pfam00112, Peptidase_C1, Papain family cysteine protease.
Aco021005.1	408	403	372	187	312	S	pfam00112, Peptidase_C1, Papain family cysteine protease.
Aco017299.1	402	399	0	2	271	S	pfam00112, Peptidase_C1, Papain family cysteine protease.
Aco024880.1	395	369	82	12	259	S	cd05472, cnd41_like, Chloroplast Nucleoids DNA-binding Protease,
Aco007211.1	382	314	23	3	233	Y	pfam00656, Peptidase_C14, Caspase domain.
Aco018123.1	365	433	378	468	406	Y	TIGR01249, Putative_proline_iminopeptidase, proline iminopeptidase,
Aco018849.1	322	355	0	72	263	-	cd08012, M20_ArgE-related, M20 Peptidases with similarity to
Aco019014.1	296	298	378	876	859	-	pfam00112, Peptidase_C1, Papain family cysteine protease.
Aco028298.1	272	262	1	42	218	-	cd08012, M20_ArgE-related, M20 Peptidases with similarity to
Aco017558.1	262	244	2	0	169	S	pfam00112, Peptidase_C1, Papain family cysteine protease.
Aco002588.1	218	224	345	239	215	Y	cd00501, Peptidase_C15, Pyroglutamyl peptidase (PGP) type I, also
Aco013996.1	206	220	34	3	146	S	pfam00450, Peptidase_S10, Serine carboxypeptidase.
Aco003873.1	197	206	118	120	164	M	COG0612, PqqL, Predicted Zn-dependent peptidases [General function
Aco021295.1	183	187	14	23	133	S	pfam00450, Peptidase_S10, Serine carboxypeptidase.

Table 6 (continued)

ID	Leaf	Flower	Root	Fruit1_5	Fruit6_8	Location	Conserved domain
Aco000819.1	178	173	0	3	117	S	cd05472, cnd41_like, Chloroplast Nucleoids DNA-binding Protease,
Aco008566.1	171	159	206	129	135	-	TIGR02227, Inactive_signal_peptidase IA., signal peptidase I,
Aco012022.1	164	156	0	0	107	S	cd04852, Peptidases_S8_3, Peptidase S8 family domain,
Aco012896.1	159	163	61	28	115	Y	cd08012, M20_ArgE-related, M20 Peptidases with similarity to
Aco005969.1	148	151	110	121	128	M	COG0612, PqqL, Predicted Zn-dependent peptidases [General function
Aco012854.1	143	142	45	33	104	Y	pfam05903, Peptidase_C97, PPPDE putative peptidase domain. The
Aco005846.1	143	149	604	21	142	S	pfam01650, Peptidase_C13, Peptidase C13 family. Members of this
Aco004358.1	137	131	3	39	92	Y	pfam00112, Peptidase_C1, Papain family cysteine protease.
Aco004359.1	133	123	92	12	85	S	pfam00112, Peptidase_C1, Papain family cysteine protease.
Aco019689.1	126	130	83	84	108	-	PRK00913, PRK00913, multifunctional aminopeptidase A; Provisional.
Aco010227.1	126	148	103	154	123	Y	cd09600, M1_APN_1, Peptidase M1 family containing Aminopeptidase N.
Aco014573.1	123	113	21	10	83	N	pfam05903, Peptidase_C97, PPPDE putative peptidase domain. The
Aco003144.1	122	132	30	28	92	P	PLN00049, PLN00049, carboxyl-terminal processing protease;
Aco009319.1	121	123	165	130	102	T	pfam04258, Peptidase_A22B, Signal peptide peptidase. The members
Aco011810.1	120	134	84	37	92	C	COG1505, COG1505, Serine proteases of the peptidase family S9A
Aco012864.1	117	121	145	143	134	-	cd09616, Peptidase_C12_UCH_L1_L3, Cysteine peptidase C12 containing
Aco027767.1	111	124	366	222	216	T	cd02620, Peptidase_C1A_CathepsinB, Cathepsin B group; composed of
Aco012310.1	110	123	78	145	116	Y	cd01085, APP, X-Prolyl Aminopeptidase 2. E.C. 3.4.11.9. Also known

Table 6 (continued)

ID	Leaf	Flower	Root	Fruit1_5	Fruit6_8	Location	Conserved domain
Aco018144.1	110	108	12	6	74	S	pfam00112, Peptidase_C1, Papain family cysteine protease.
Aco016160.1	109	120	15	30	85	Y	COG0542, clpA, ATP-binding subunits of Clp protease and DnaK/DnaJ
Aco011447.1	107	117	15	92	77	-	cd04852, Peptidases_S8_3, Peptidase S8 family domain,
Aco010611.1	106	118	33	2	75	S	pfam00450, Peptidase_S10, Serine carboxypeptidase.
Aco006954.1	101	101	3	1	68	N	pfam05903, Peptidase_C97, PPPDE putative peptidase domain. The
Aco028297.1	101	109	3	27	75	-	cd08012, M20_ArgE-related, M20 Peptidases with similarity to
Aco012083.1	97	102	49	70	85	C	pfam00450, Peptidase_S10, Serine carboxypeptidase.
Aco005269.1	96	103	45	37	77	S	pfam00112, Peptidase_C1, Papain family cysteine protease.
Aco008796.1	95	103	93	56	74	T	pfam04258, Peptidase_A22B, Signal peptide peptidase. The members
Aco025188.1	1	1	483	0	1	S	pfam00112, Peptidase_C1, Papain family cysteine protease.
Aco001368.1	1	0	411	0	0	S	pfam00112, Peptidase_C1, Papain family cysteine protease.
Aco006487.1	5	5	121	198	21	S	cd05472, cnd41_like, Chloroplast Nucleoids DNA-binding Protease,
Aco014769.1	27	25	118	13	20	-	PLN03146, PLN03146, aspartyl protease family protein; Provisional.
Aco008415.1	27	22	116	151	85	N	pfam02338, OTU, OTU-like cysteine protease. This family is
Aco005786.1	64	69	113	79	65	C	cd02657, Peptidase_C19A, A subfamily of Peptidase C19. Peptidase
Aco026097.1	84	76	101	57	76	N	pfam00656, Peptidase_C14, Caspase domain.
Aco007611.1	56	60	101	80	58	Y	PLN03158, PLN03158, methionine aminopeptidase; Provisional.
Aco004901.1	92	71	62	294	112	S	cd04852, Peptidases_S8_3, Peptidase S8 family domain,
Aco012554.1	70	77	88	190	84	T	cd02125, PA_VSR, PA_VSR: Protease-associated (PA) domain-containing
Aco010208.1	80	91	60	161	81	N	pfam02338, OTU, OTU-like cysteine protease. This family is

Table 6 (continued)

ID	Leaf	Flower	Root	Fruit1_5	Fruit6_8	Location	Conserved domain
Aco004360.1	73	68	15	160	91	S	pfam00112, Peptidase_C1, Papain family cysteine protease.
Aco013115.1	60	65	87	129	76	Y	TIGR00763, Lon_protease, endopeptidase La. This protein, the
Aco023270.1	56	59	73	106	81	P	pfam06703, SPC25, Microsomal signal peptidase 25 kDa subunit
Aco018628.1	36	36	43	104	54	C	cd07017, S14_ClpP_2, Caseinolytic protease (ClpP) is an
Aco011236.1	54	54	67	104	68	S	cd02123, PA_C_RZF_like, PA_C-RZF_like: Protease-associated (PA)

The list contains the protease coding genes having >100 FPKM. The complete list is showed as Supplementary Table 7. The numbers are the normalized FPKM (fragments per kilobase exon per million reads mapped) values of each sample. The value in bold is the tissue having gene expression up-regulated relative to the tissue with a number in italic and underlined

Subcellular location notation *C* chloroplast, *E* ER, *G* Golgi, *K* cytoskeleton, *M* mitochondrial, *N* nuclear, *O* Peroxisome, *P* plasma membrane, *S* secreted, *T* membrane, *V* Vacuole, *Y* cytoplasm

Computational Prediction Methods for Protein Subcellular Locations

For protein subcellular location prediction we used a computational protocol that was developed based on the evaluation of the prediction accuracy in plants (Min 2010; Meinken and Min 2012; Lum et al. 2013, 2014). In brief, the protein subcellular locations and features were computationally predicted using multiple software tools, including SignalP 4.0, TargetP, Phobius, WoLF PSORT, TMHMM, and PS-Scan (Lum and Min 2013). The software tools were installed on a local Linux system for data processing. The detailed input, output, and the commands for how to run them were explained and summarized by Lum and Min (2013). In brief, SignalP 4.0, Phobius, and TargetP were used for secretory signal peptide prediction (Petersen et al. 2011; Käll et al. 2004; Emanuelsson et al. 2007). TargetP also predicts the presence of chloroplast transit peptide (cTP) or mitochondrial targeting peptide (mTP) in the N-terminus (Emanuelsson et al. 2007). TMHMM was used to predict the presence and topology of transmembrane helices (Krogh et al. 2001). PS-Scan was used to scan the PROSITE database for removing ER targeting proteins (Prosite: PS00014) (Sigrist et al. 2010). WoLF PSORT predicts multiple subcellular locations including chloroplast, 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.

Based on our previous evaluation and experiences, the secreted proteins were assigned to entries that are predicted to have a secretory signal peptide by at least two of the four predictors - SignalP 4.0, Phobius, TargetP and WoLF PSORT - and that do not have a transmembrane domain or an ER retention peptide (Min 2010; Lum et al. 2014). As we attempted to find the maximum number of secreted proteins without sacrifice too much in specificity, thus the secretome dataset includes the predicted highly likely secreted proteins (a signal peptide predicted by at least 3 of 4 predictors) and likely secreted proteins (a signal peptide predicted by 2 of 4 predictors), as listed in the category of the PlantSecKB database (Lum et al. 2014). The method for the above secretome prediction had a sensitivity of 79.1 %, a specificity of 98.6 %, and a Mathews' correlation coefficient (MCC) of 0.73, which were much better than using a single predictor alone (Min 2010; Meinken and Min 2012; Lum et al. 2014). Chloroplast proteins were predicted to be chloroplast located by both TargetP and WoLF PSORT, however, mitochondrial proteins were assigned based on the prediction of WoLF PSORT only. The method for the organelle protein prediction was slightly different from our previous adopted method of using TargetP only (Lum et al. 2014). The prediction accuracy for chloroplast proteins was 96.1 % in specificity and 22.7 % in sensitivity, and for mitochondrial proteins was 96.2 % in specificity and 33.3 % in sensitivity. ER proteins include proteins predicted to be located by WoLF PSORT and proteins to contain a signal peptide by SignalP 4.0 and an ER target signal (Prosite: PS00014) by PS-Scan. Other subcellular locations

including cytosol (cytoplasm), cytoskeleton, Golgi apparatus, lysosome, nucleus, peroxisome, plasma membrane and vacuole were predicted by WoLF PSORT. The overall accuracies for other subcellular protein locations were evaluated previously (Lum et al. 2014).

Conserved Domain and Gene Ontology (GO) Classification

For the predicted secretome datasets in each species, we mapped them to Gene Ontology (GO) using BLASTP against the UniProt Swiss-Prot dataset with an E-value cutoff 1e-5. The GO IDs were retrieved from the UniProt ID mapping data and then obtained a summary table for each main category including biological process, molecular function, and cellular component using the GO SlimViewer (http://www.agbase.msstate.edu/cgi-bin/tools/goslimviewer_select.pl) (McCarthy et al. 2006). The proteome datasets were further analyzed functionally by searching the conserved domain database using rpsBLAST with a cutoff E-value of 1e-5. The conserved domain database contains multiple datasets including PFam, SMART, the COGs collection, TIGRFAMs, NCBI's in-house data curation effort, etc. (Marchler-Bauer et al. 2015).

cDNA Library Construction, Sequencing and RNA-seq Expression Analysis

Pineapple leaf, root, and flower tissues were collected from var. F153 plants and fruit tissues were harvested from hybrid MD-2 for RNA extraction and transcriptome sequencing. RNA was extracted from leaf, root and flower tissues using Qiagen RNeasy Plant Mini Kit (Qiagen, #74904), following manufacturer's protocols. For pineapple fruit RNA, it was extracted using salt buffer followed with lithium chloride and isopropanol precipitation. Then, DNA was removed with DNA-free™ DNA Removal Kit (Life Technologies, #AM1906M). Single indexed RNA-seq library was constructed using Illumina TruSeq stranded RNA Sample Preparation Kit (Illumina, #RS-122-2001) and then sequenced by Illumina HiSeq2500 in single end 100 nt mode.

For expression profiling at gene level, the trimmed single end reads of each sample were aligned to repeat-masked pineapple assembly version 3 using TopHat v2.0.9 default setting (Trapnell et al. 2012). The normalized FPKM (fragments per kilobase exon per million read mapped) value of each sample were estimated by Cufflinks v2.2.1, followed by Cuffnorm v2.2.1 using default setting with pineapple gene model annotation provided (-g option). We combined the five stages (stage 1 to 5) of immature fruit into one sample as young fruit and the late 3 stages (stage 6 to 8) into one sample as ripening fruit, and pairwise compared the five tissues. We further identified genes down-regulated or up-regulated based on at least two fold changes of the log₂ values of FPKM, however, genes

not having more than 10 FPKM in at least one tissue were treated as lowly expressed genes and not used for pairwise comparison.

Data and Tool Access

The data described can be accessed and downloaded through the web user interface at PlantSecKB (<http://proteomics.yzu.edu/secretomes/plant.php>) and some intermediate data and supplementary files can be downloaded (http://bioinformatics.yzu.edu/publication/data/Pineapple/subloc_proteases/). The interface provides various utilities for searching proteins obtained from UniProt database and proteins predicted in sacred lotus and pineapple genome (Lum et al. 2013, 2014). All plant proteins can be searched using UniProt accession number (AC) or protein ID for pineapple and sacred lotus, key word(s), or species. A BLAST utility can be accessed through a link on the interface to search plant all proteins or secretomes. It should be noted that the subcellular proteome data downloaded from the PlantSecKB database may be slightly different from the data described in the paper as the data in the database have some redundant collections.

Acknowledgments The work was supported by the University of Illinois at Urbana-Champaign to RM and Youngstown State University (YSU) Research Council 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
- Amid A, Ismail NA, Yusof F, Salleh HM (2011) Expression, purification, and characterization of a recombinant stem bromelain from *Ananas comosus*. *Process Biochem* 46:2232–2239
- Arshad ZIM, Amid A, Yusof F et al (2014) Bromelain: an overview of industrial application and purification strategies. *Appl Microbiol Biotechnol* 98:7283–7297
- Barrett AJ, McDonald JK (1986) Nomenclature: protease, proteinase and peptidase. *Biochem J* 237:935
- Bartholomew DP, Kadzimin SB (1977) Pineapple. In: Alvin PT, Kozeowski TT (eds) *Ecophysiology of tropical crops*. Academic, New York, NY, pp 113–156
- Bartholomew DP, Malézieux EP (1994) Pineapple. In: Schaffer B, Andersen PC (eds) *Handbook of environmental physiology of fruit crops*, vol 2. CRC Press, Boca Raton, pp 243–291
- Carter CE, Marriage H, Goodenough PW (2000) Mutagenesis and kinetic studies of a plant cysteine proteinase with an unusual arrangement of acidic amino acids in and around the active site. *Biochemistry* 39: 11005–11013
- de Lencastre Novaes LC, Jozala AF, Lopes AM et al (2016) Stability, purification, and applications of bromelain: a review. *Biotechnol Prog* 32:5–13. doi:10.1002/btpr.2190
- 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

- George S, Bhasker S, Madhav H, Nair A, Chinnamma M (2014) Functional characterization of recombinant bromelain of *Ananas comosus* expressed in a eukaryotic system. *Mol Biotechnol* 56:166–174
- González-Rábade N, Badillo-Corona JA, Aranda-Barradas JS, del Carmen O-SM (2011) Production of plant proteases in vivo and in vitro—a review. *Biotechnol Adv* 29:983–996
- Habib H, Fazili KM (2007) Plant protease inhibitors: a defense strategy in plants. *Biotechnol Mol Biol Rev* 2:68–85
- Heinicke RM, Gortner WA (1957) Stem bromelain - a new protease preparation from pineapple plants. *Econ Bot* 11:225–234
- Horton P, Park K-J, Obayashi T et al (2007) WoLF PSORT: protein localization predictor. *Nucleic Acids Res* 35:585–587
- 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
- Jung YJ, Choi CS, Park JH et al (2008) Overexpression of the pineapple fruit bromelain gene (BAA) in transgenic Chinese cabbage (*Brassica rapa*) results in enhanced resistance to bacterial soft rot. *Electron J Biotechnol* 11:71–79
- Käll L, Krogh A, Sonnhammer EL (2004) A combined transmembrane topology and signal peptide prediction method. *J Mol Biol* 338:1027–1036
- Krogh A, Larsson B, von Heijne G, Sonnhammer ELL (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol* 305:567–580
- Lee KL, Albee KL, Bernasconi RJ, Edmunds T (1997) Complete amino acid sequence of ananain and a comparison with stem bromelain and other plant cysteine proteases. *Biochem J* 327(Pt 1):199–202
- Lopez-Casado G, Urbanowicz BR, Damasceno CM, Rose JK (2008) Plant glycosyl hydrolases and biofuels: a natural marriage. *Curr Opin Plant Biol* 11:329–337
- López-García B, Hernández M, Segundo BS (2012) Bromelain, a cysteine protease from pineapple (*Ananas comosus*) stem, is an inhibitor of fungal plant pathogens. *Lett Appl Microbiol* 55:62–67
- Lum G, Min XJ (2011) Plant secretomes: current status and future perspectives. *Plant Omics* 4:114–119
- Lum G, Min XJ (2013) Bioinformatic protocols and the knowledge-base for secretomes in fungi. In: Gupta VK, Tuohy MG, Ayyachamy M, Turner KM, O'Donovan A (eds) *Laboratory protocols in fungal biology: current methods in fungal biology*, Springer pp 545–557.
- Lum G, Vanburen R, Ming R, Min XJ (2013) Secretome prediction and analysis in sacred lotus (*Nelumbo nucifera* Gaertn.). *Trop Plant Biol* 6:131–137
- Lum G, Meinken J, Orr J, Frazier S, Min XJ (2014) PlantSecKB: the plant secretome and subcellular proteome knowledgebase. *Computat Mol Biol* 4(1):1–17
- Marchler-Bauer A, Derbyshire MK, Gonzales NR et al (2015) CDD: NCBI's conserved domain database. *Nucleic Acids Res* 43:222–226
- McCarthy FM, Wang N, Magee GB et al (2006) AgBase: a functional genomics resource for agriculture. *BMC Genomics* 7:229
- Meinken J, Min XJ (2012) Computational prediction of protein subcellular locations in eukaryotes: an experience report. *Comput Mol Biol* 2(1):1–7. doi:10.5376/cmb.2012.02.0001
- Meiser P, Xu Z, Kirsch G, Jacob C (2014) Systemic enzyme therapy: fact or fiction? a review with focus on bromelains, proteolytic enzymes from the pineapple plant. In: *Recent Advances in Redox Active Plant and Microbial Products*. Springer Netherlands. p 449–467
- Min XJ (2010) Evaluation of computational methods for secreted protein prediction in different eukaryotes. *J Proteomics Bioinfo* 4:143–147
- Ming R, VanBuren R, Wai CM et al (2015) The pineapple genome and the evolution of CAM photosynthesis. *Nat Genet* 47:1435–1442. doi:10.1038/ng.3435
- Neuteboom LW, Matsumoto KO, Christopher DA (2009) An extended AE-rich N-terminal trunk in secreted pineapple cystatin enhances inhibition of fruit bromelain and is posttranslationally removed during ripening. *Plant Physiol* 151:515–527
- Pavan R, Jain S, Kumar A (2012) Properties and therapeutic application of bromelain: a review. *Biotechnol Res Int*. doi:10.1155/2012/976203
- Petersen TN, Brunak S, von Heijne G, Nielsen H (2011) SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat Methods* 8:785–786
- Raimbault AK, Zuily-Fodil Y, Soler A et al (2013a) The expression patterns of bromelain and AcCYS1 correlate with blackheart resistance in pineapple fruits submitted to postharvest chilling stress. *J Plant Physiol* 170:1442–1426
- Raimbault AK, Zuily-Fodil Y, Soler A, de Carvalho MHC (2013b) A novel aspartic acid protease gene from pineapple fruit (*Ananas comosus*): cloning, characterization and relation to postharvest chilling stress resistance. *J Plant Physiol* 170:1536–1540
- Sigrist CJA, Cerutti L, de Casro E et al (2010) PROSITE, a protein domain database for functional characterization and annotation. *Nucleic Acids Res* 38:161–166
- Taussig SJ, Batkin S (1988) Bromelain, the enzyme complex of pineapple (*Ananas comosus*) and its clinical application: an update. *J Ethnopharmacol* 22:191–203
- 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
- Trapnell C, Roberts A, Goff L et al (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc* 7:562–578
- UniProt Consortium (2015) UniProt: a hub for protein information. *Nucleic Acids Res* 43:204–212. doi:10.1093/nar/gku989
- van der Biezen EA, Jones JD (1998) The NB-ARC domain: a novel signalling motif shared by plant resistance gene products and regulators of cell death in animals. *Curr Biol* 8:226–227
- van der Hoorn RA (2008) Plant proteases: from phenotypes to molecular mechanisms. *Annu Rev Plant Biol* 59:191–223
- van Wijk KJ (2015) Protein maturation and proteolysis in plant plastids, mitochondria, and peroxisomes. *Annu Rev Plant Biol* 66:75–111
- Wai CM, Powell B, Ming R, Min XJ (2016) Analysis of alternative splicing landscape in pineapple (*Ananas comosus*). *Tropical Plant Biol*. doi:10.1007/s12042-016-9168-1
- Wang W, Zhang L, Guo N et al (2014) Functional properties of a cysteine proteinase from pineapple fruit with improved resistance to fungal pathogens in *Arabidopsis thaliana*. *Molecules* 19:2374–2389