研究报告
Research Report

PlantSecKB:植物分泌蛋白质组亚细胞蛋白质组知识库

Gengkon, John, Jessica, Stephanie, Xiang

1 计算机科学与信息系, 扬斯敦州立大学, 扬斯敦, OH 44555, 美国

2 生物科学系, 扬斯敦州立大学, 扬斯敦, OH 44555, 美国

3 应用化学生物学中心, 扬斯敦州立大学, 扬斯敦, OH 44555, 美国

通讯作者, xmin@ysu.edu; 作者

计算分子生物学, 2014 年, 第 3 卷, 第 8 篇

收稿日期: 2014 年 09 月 07 日 接受日期: 2014 年 09 月 07 日 发表日期: 2014 年 09 月 07 日

摘 要 蛋白质亚细胞位置的预测和处理是蛋白质功能注释的必要条件。我们开发植物分泌蛋白质组和亚细胞蛋白质组知识库(PlantSecKB)方便植物研究界获取和处理植物蛋白，特别是分泌蛋白的亚细胞位置。从 UniProtKB 数据库检索植物蛋白序列所有可用的植物蛋白质数据构建数据库，由 PlantGDB 项目组装的 EST 数据库进行预测。数据库包含从三个来源收集的信息：(1)在 UniProtKB 中创建或计算预测的亚细胞位置；(2)亚细胞位置和特征由八个计算工具预测；(3)分泌蛋白从最近的文献获取。亚细胞位置的类别包括分泌蛋白，线粒体，叶绿体，细胞质，细胞骨架，内质网，高尔基体，溶酶体，过氧化物酶体，细胞核，液泡和质膜。数据可以通过使用 UniProt 登录号、ID、GenBank GI 或 RefSeq 登录号、基因名称和关键词来搜索。可以搜索物种特异性分泌蛋白和亚细胞蛋白质组学，并将其以 FASTA 文件下载。BLAST 允许用户基于蛋白质序列搜索数据库，也支持植物蛋白亚细胞位置的管理。一项主要的分析显示单子叶植物和双子叶植物具有相似比例，单子叶植物在线粒体和叶绿体膜(包括膜和非膜)中分布着显著的更高比例的蛋白质，而双子叶植物具有显著更多的蛋白质分布在胞质溶胶和细胞核中。该数据库旨在促进植物蛋白质研究，可在 http://proteomics.ysu.edu/secretomes/plant.php 获取信息。

关键词 计算预测；表达序列标签；植物；分泌蛋白；分泌蛋白组；信号肽；亚细胞定位；亚细胞蛋白质组

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 forApplied Chemical Biology, Youngstown State University, Youngstown, OH 44555, USA

Corresponding author, xmin@ysu.edu; Authors

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

植物是生物质的主要生产者，包括碳水化合物，蛋白质，脂质，纤维素和其他。植物蛋白质包括酶、调节和结构蛋白质，在调节植物生长和发育中发挥重要的生物学作用。植物蛋白在细胞内合成然后转运到不同的亚细胞位置，包括细胞外空间或基质，以发挥它们的生物学功能。这个过程通常被称为蛋白质分选和靶向(Foresti and Denecke, 2008; Rose and Lee, 2010)。植物细胞含有细胞壁，质膜，叶绿体，线粒体，大液泡，细胞核，内质网(ER)，高尔基体，过氧化物酶体，胞质溶胶等。膜蛋白可以嵌入或连接到质膜、细胞器膜或内膜系统。

真核生物中蛋白质亚细胞位置的鉴定和分析是注释蛋白质组的重要课题之一。在植物物种中，分泌到细胞外空间或基质(包括细胞壁)的蛋白质统称为“分泌蛋白”(Agrawal et al., 2010; Lum and Min, 2011a)。这一术语最初由 Tjalsma 等(2000)提出，表示由分泌途径加工的枯草芽孢杆菌完整的蛋白质组，其包括分泌到细胞外空间以及参与途径的蛋白质。但实际上往往有所限制，比如这项工作中仅表示蛋白质组包括细胞壁蛋白分泌的胞外部分(e.g., Greenbaum et al., 2001; Hathout, 2007; Bouws et al., 2008; Agrawal et al., 2010; Lum and Min, 2011b)。植物分泌组蛋白主要由细胞壁蛋白、参与细胞壁代谢的蛋白质以及涉及病原体防御的胞外酶和信号分子组成(Isaacson and Rose, 2006; Kamoun, 2009; Lum and Min, 2011a)。分泌酶，特别是水解酶如 α-淀粉酶和 α-葡糖苷酶，已经使用发芽大麦种子作为模型系统进行了充分研究。这些水解酶在糊粉层中合成并分泌到胚乳中以分解淀粉和其他储存物(Ranki and Sopanen, 1984; Jones and Robinson, 1989; Finnie et al., 2011 for review)。最近，蛋白质组学分析技术以及拟南芥和水稻基因组完整的测序的进展产生许多分泌的蛋白质，包括检测到的细胞壁蛋白质组(Boudart et al., 2007; Agrawal et al., 2010; Lum and Min, 2011a)，这些鉴定的分泌蛋白主要由拟南芥中的细胞壁蛋白(see Jamet et al., 2008 for review)和诸如参与病原体防御的酶 GLP1(Oh et al., 2005)组成。使用叶或种子细胞悬浮培养，分泌蛋白用 2D 凝胶电泳结合水稻、苜蓿和高粱的液相色谱质谱分析鉴定(Jung et al., 2008; Kusumawati et al., 2008; Cho et al., 2009; Ngara and Ndimba, 2011)。从无菌生长的水稻和拟南芥的幼苗根部分泌物中也鉴定出了大量的分泌蛋白(Shinano et al., 2011; De-la-Pena et al., 2010)。最近用于植物分泌组织研究的实验系统，分析技术和相关的生物信息学工具有了全面的综述(Agrawal et al., 2010; Meinken and Min, 2012; Alexandersson et al., 2013; Kraus et al., 2013; Caccia et al., 2013)。

经典的真核分泌蛋白在 N-末端含有将蛋白质导向粗面内质网以完成蛋白质合成然后将其运输到高尔基复合体以获得目标蛋白的分泌信号肽(von Heijne, 1990)。信号肽通常为 15~30 个氨基酸的长度，经常会在通过内膜系统移位期间被切除。经典分泌的蛋白质可以相对精确地计算预测(Min, 2010)。最近我们分析了 UniProtKB / Swiss-Prot 数据集中所有人为选择和注释的植物分泌蛋白，发现 87％可以被使用的三个预测器预测到含有信号肽(Lum and Min, 2011a)。通过使用新版本的 SignalP (SignalP 4.0)结合其他工具，包括用于鉴定跨膜蛋白的 TMHMM、用于鉴定内质网腔蛋白的 PS-Scan，分泌物预测的准确性可以进一步提高。(Min, 2010; Melhem et al., 2013)。

随着测序技术的改进和测序成本的降低，越来越多的植物物种的基因组被完全测序。目前有 32 个具有完整或原始基因组序列的陆地植物以及 73 种陆地植物正在进行基因组测序(http://www.ncbi.nlm.nih.gov/genomes/static/gpstat.html)。在植物中还存在组装的表达序列标签(EST)数据，用于鉴定超过 200 种物种中编码分泌蛋白的潜在基因(PlantGDB, http://www.plantgdb.org/prj/ESTCluster/) (Duvick et al., 2008)。

作为基因组测序的结果，可用蛋白质序列的数量迅速增加。除了经典的分泌蛋白之外，还在植物

中鉴定了大量无铅，非经典，即不具有分泌信号肽的分泌蛋白(LSP) (Jung et al., 2008; Agrawal et al., 2010; Ding et al., 2012 for review)。这些蛋白质没有在 UniProtKB 中创建，因此，需要有一个中心知识库为植物研究界提供植物蛋白亚细胞位置以获取可用的信息并为新表征的蛋白质沉积实验证据。为了提供这样的中心植物分泌物相关资源门户，我们开发了植物分泌蛋白组和亚细胞蛋白质组知识库(PlantSecKB)(http://proteomics.ysu.edu/secretomes/plant.html)，包括来自植物蛋白酶的预测、人为选择的蛋白质亚细胞位置以及来自植物 EST 数据的预测蛋白质。我们的重点是植物分泌蛋白，但是还提供了位于其他亚细胞位置的蛋白质信息。可以通过数据库界面访问用于植物蛋白亚细胞位置的人为选择管理的工具。

1 数据库构造方法

1.1 数据采集

PlantSecKB 主要用从两个来源获得的序列数据构建：从 UniProtKB 提取的植物蛋白序列(2013-04 发布) (http://www.uniprot.org/)，从 PlantGDB 编译的 EST 数据组合中预测的蛋白质序列(http://www.plantgdb.org/prj/ESTCluster/)。最近从圣莲(Nelumbo nucifera Gaertn.)中预测的基因组蛋白也被整合到该数据库中(Ming et al., 2013; Lum et al., 2013)。EST 数据中的蛋白质序列使用具有针对 UniProt / Swiss-Prot 数据库的 BLASTX 输入的 OrfPredictor 工具预测(http://proteomics.ysu.edu/tools/OrfPredictor.html)，TargetIdentifier (http://proteomics.ysu.edu/tools/TargetIdentifier.html)用于检查 EST 是否是全长的(Min et al., 2005a, 2005b)。

1.2 用于预测蛋白质亚细胞位置的计算方法

本研究中使用的软件工具包括 SignalP 3.0 和 4.0、TargetP、Phobius、WoLF PSORT、TMHMM、PS-Scan 和 FragAnchor。这些工具和参考的相关网站链接可以在我们的网站中找到(http://proteomics.ysu.edu/tools/subcell.html)。除了 Frag Anchor，我们使用安装在本地 Linux 系统上的独立工具进行数据处理。有关运行它们的命令可以在每个下载的软件包中的“readme”页面中找到，Lum 和 Min 进行了总结(2013)。简言之，SignalP 4.0 用于分泌信号肽的预测(Petersen et al., 2011)。也包括了来自 SignalP 3.0 的预测信息(Bendtsen et al., 2004b)因为其提供比 SignalP 4.0 更准确的切割位点预测(Petersen et al., 2011)。Phobius 是一个组合的信号肽和跨膜拓扑预测(Käll et al., 2007)。靶标预测用于预测在 N-端存在例如信号肽(SP)、叶绿体转运肽(cTP) 或线粒体靶向肽(mTP)的信号序列(Emanuelsson et al., 2000; Emanuelsson et al., 2007)。TMHMM 使用隐马尔可夫模型(HMM)来预测跨膜螺旋的存在和拓扑及其对膜的定向(进/出) (Krogh et al., 2001)。PS 扫描用于扫描 PROSITE 数据库(http://www.expasy.org/tools/scanprosite/)以去除 ER 靶向蛋白(Prosite: PS00014) (de Castro et al., 2006; Sigrist et al., 2010)。FragAnchor 用来鉴定通过 SignalP 4.0 预测含有信号肽的蛋白质中的糖基磷脂酰肌醇(GPI)锚定蛋白(GAP) (Poisson et al., 2007)。WoLF PSORT 可以预测多种亚细胞位置，包括胆石、胞质溶胶、细胞骨架、ER、胞外(分泌的)、高尔基体、溶酶体、线粒体、核、过氧化物酶体、质膜和液泡膜(Horton et al., 2007)。真核生物或植物的默认参数(如果可用)用于所有程序。我们先前的评估发现，用于植物分泌组织预测的 WoLF PSORT 会由于预测灵敏度的显著降低而导致准确度降低(Min, 2010)。因此，其不用于分泌物预测，仅用于预测一些其它亚细胞位置。

关于分配蛋白质的亚细胞位置，UniProtKB 注释的亚细胞位置和我们的人为选择优先于计算预测，因此，只对没有注释的亚细胞位置的蛋白质进行亚细胞位置的计算分配。所有工具产生的信息仍然可用于所有植物蛋白质。一些蛋白质可以具有一个以上亚细胞位置，以下标准用于蛋白质亚细胞位置的计算分类：

膜蛋白：通过 TMHMM 预测含有一个或多个跨膜结构域的蛋白质。但是，如果仅预测到一个跨膜结构域，位于 N-末端 70 个氨基酸内，并且被 SignalP4.0 预测为信号肽，则该蛋白质不被视为膜蛋白。

叶绿体蛋白：被 TargetP 预测为“C”(对于叶绿体)用于亚细胞定位的蛋白质。如果它也被归类为膜蛋白，则其被进一步分类为叶绿体膜蛋白。

线粒体蛋白：由 TargetP 预测为亚型细胞定位的“M”(线粒体)的蛋白质。如果它也被分类为膜蛋

白，则进一步分类为线粒体膜蛋白。

ER 蛋白：通过 SignalP 4.0 预测含有信号肽的蛋白质和通过 PS 扫描含有 EHR 靶标信号(Prosite：PS00014)。

完整的分泌组蛋白：来自一个物种的所有的分泌蛋白。通过所有三个预测 SignalP 4.0、Phobius 和 TargetP 预测具有分泌信号肽并且不被分类为任何上述类别的蛋白质，未分类为任何上述类别并且被一个或两个预测器预测为的信号肽的蛋白质被认为“弱可能分泌的”或“可能分泌的”，因为我们先前的评估揭示信号肽在一些注释的分泌蛋白质中只能被一个或两个预测器检测到(Lum and Min, 2011a)。联合使用三个预测器可以增加分泌物预测的特异性，提高预测准确性(Min, 2010; Melhem et al., 2013)。所有人为选择的卷曲分泌蛋白和细胞外蛋白质包括在完整分泌蛋白中。

卷曲分泌蛋白：该类别包括在来自 UniProtKB / Swiss-Prot 数据集的“审查”的亚细胞位置中注释为“分泌型”或“细胞外”或“细胞壁”的蛋白质，它还包括我们从最近的文献中人工手机的分泌蛋白。

GPI 锚定蛋白：被 Frag Anchor 预测具有 GPI 锚含信号肽的蛋白质进一步分类为 GPI 锚定蛋白。预测具有信号肽和 GPI 锚的蛋白质序列可附着于质膜的外部小叶或分泌成为细胞壁的组分。这些蛋白质参与信号传导、粘附、应激反应、细胞壁重塑或在生长和发育中发挥其他作用(Borner et al., 2002; Borner et al., 2003; Gillmor et al., 2005; Simpson et al., 2009)。

其他亚细胞位置的蛋白质：由 WoLF PSORT 预测的其它亚细胞位置包括细胞质(细胞质)、细胞骨架、高尔基体、溶酶体、细胞核、过氧化物酶体、质膜和液泡。

1.3 蛋白质亚细胞位置的计算预测精度

我们上面使用的预测方法的开发是基于我们以前对计算工具的评估(Min, 2010; Meinken and Min, 2012; Melhem et al., 2013)。为了估计我们的方法对每个亚细胞定位的预测精度，我们使用了两个数据集(表 1)。数据集 A 由 15 028 个蛋白组成，该数据集包含来自具有人为选择亚细胞位置的 UniProtKB / Swiss-Prot 数据集的蛋白质，排除具有多个亚细胞位置或标记为“片段”的蛋白质。数据集 B 由 6 908 个蛋白组成，这些蛋白在排除在亚细胞位置注释中具有术语“通过相似性”或“可能”或“预测”的条目之后从数据集 A 生成。与使用单个工具的其他方法相比，我们的方法是使用多种工具的组合，包括 SignalP 4.0、TargetP 和 Phobias 用于分泌信号肽预测、PS 扫描用于去除 ER 蛋白和用 TMHMM 用于去除膜蛋白，显著提高了分泌信号肽的预测准确性(Min, 2010; Meinken and Min, 2012)。对于数据集 A 的分泌组蛋白预测，我们的方法达到 91.1％的灵敏度，98.7％的特异性和 88.5％的 Mathews 相关系数(MCC)，对于数据集，灵敏度为 76.8％，特异性为 98.9％，MCC 为 74.5％。这比单独使用 WoLF PSORT 或 MultiLoc 好得多(Meinken and Min, 2012)，因此，分泌蛋白的预测是相对可靠的。预测其他亚细胞位置的准确性仍需要改进。

1.4 人工管理和团体注释

基于已发表的实验证据，PlantSecKB 为植物蛋白亚细胞位置的管理提供支持。一个为团队服务的提交工具被开发用来提供蛋白质的亚细胞位置注释以及支持它的注释文献来源。经过我们的验证，这些数据也被合并到数据库中。目前，根据已发表的实验数据，我们从水稻(Jung et al., 2008; Cho et al., 2009; Cho and Kim, 2009; Chen et la., 2009; Zhang et al., 2009; Shinano et al., 2011)，拟南芥(De-la-Pena et al., 2010)和高粱(Ngara et al., 2011)中人为选择了 736 个分泌蛋白。人为选择是一个持续的过程，因此，来自我们和其他团队的更多的分泌蛋白将被人为选择并整合到数据库中。来自计算预测、UniProtKB 注释和人为选择的信息被集成并显示在注释页面上(图 1)。注释的条目被链接到所使用的工具、UniProtKB、RefSeq 数据库和国家生物技术信息中心(NCBI)的 PubMed 上。

2 数据库内容和工具概述

2.1 数据和工具访问

PlantSecKB 通过数据库网页 http://proteomics.ysu.edu/secretomes/plant.php 界面访问。该界面提供了用于搜索从 UniProtKB 获取的蛋白质的各种实用程序，指向 BLAST 的链接、EST 数据搜索页面和注释页面(图 1)。可以使用 UniProt 登录号(AC)或 ID、基因名称、蛋白质功能或物种的关键词来搜索 UniProt 包含的所有植物蛋白。亚蛋白质组包括卷曲分泌蛋白、完全分泌蛋白、线粒体

表 1 植物蛋白亚细胞定位预测精度的评价

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

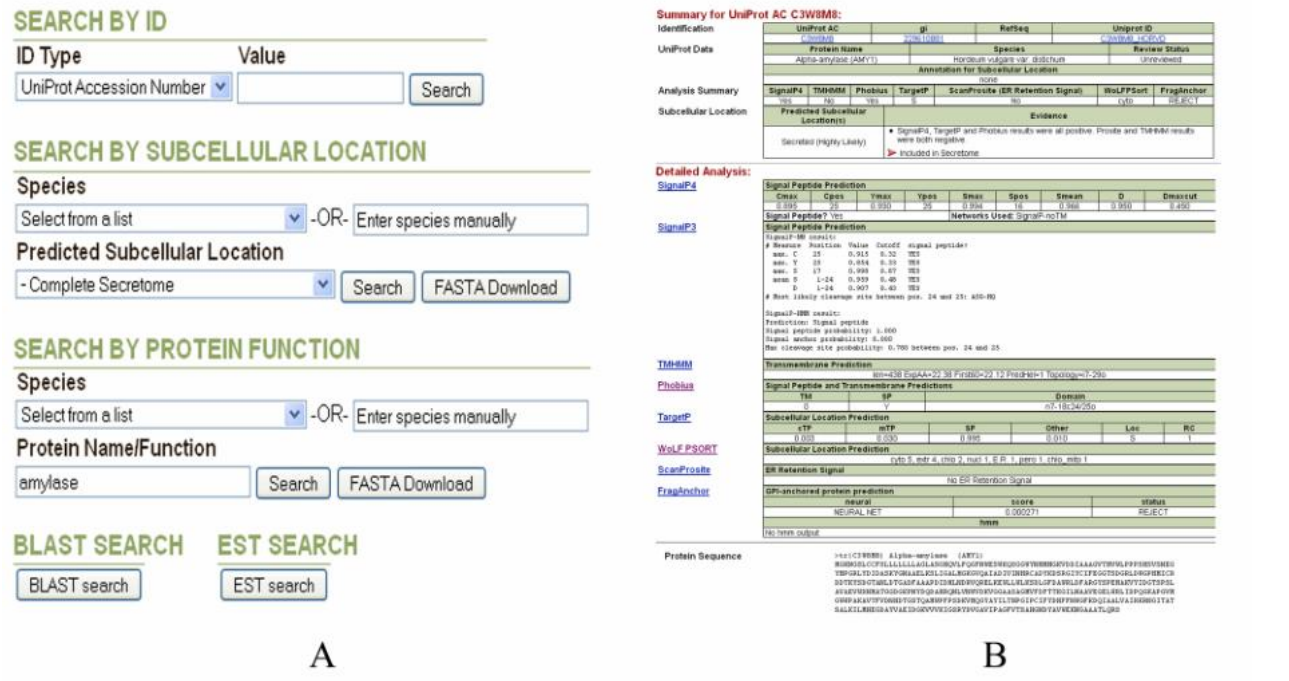

图 1 PlantS eKB 用户界面和注释页面概述

Figure 1 Overview of the PlantSecKB user interface and annotation page

膜蛋白、OR 蛋白以及其他可以通过从物种列表中选择具有大于 1 000 个蛋白质序列搜索或下载的其他物种。可以通过输入物种名称来搜索具有少于 1000 个蛋白质条目的种类。可以通过界面上的链接访问 BLAST 实用程序来搜索所有植物蛋白或分泌物，该界面还提供指向 EST 数据搜索页面的链接，可以使用 EST 标识符，关键字，种或 BLAST 进行搜索。

每个 UniProt 蛋白质的注释显示页面包含从以下三个来源获得的信息：(1)使用上述七个程序的计算方法预测的特征；(2)在 UniProtKB 中注释的亚细胞位置；(3)我们根据最新文献的实验依据进行的人

表 2 PlantSecKB 中不同植物物种的亚细胞蛋白质组学总结

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

	Total	Sec	Mt		Ch		Cyt	Cyk	Gol	Per	Nuc	Pla mem	Vac	GPI anc
			mem	non-mem	mem	non-mem								
Green algae														
Chlamydomonas reinhardtii	15120	545	435	3402	256	1605	2658	71	14	57	1641	487	131	27
Volvox carteri	14825	560	470	3019	200	1445	2478	46	15	59	2120	490	109	36
Micromonas pusilla	10811	155	242	1758	268	1854	2005	36	11	21	1557	231	83	6
Micromonas sp.	10123	184	248	1551	284	1494	2204	42	11	34	1595	324	86	38
Chlorella variabilis	9835	411	286	2013	136	910	1873	19	16	47	1213	361	80	33
Coccomyxa subellipsoidea	9799	406	210	1551	101	785	2121	36	23	27	1906	452	97	11
Guillardia theta	8029	71	274	1979	186	1234	1376	12	8	4	1037	212	42	9
Bathycoccus prasinos	7886	145	172	918	304	1646	1027	34	16	13	1877	330	54	15
Guillardia theta nucleomorph	7404	50	235	1635	72	573	1763	40	8	12	923	200	69	8
Monocots														
Oryza sativa subsp. japonica	59948	5027	1707	17419	1328	14559	15442	545	213	144	19654	5220	736	566
Zea mays	52865	3883	1161	10333	1106	9632	9812	357	135	105	10810	1895	540	372
Oryza sativa subsp. indica	40429	2646	772	6199	561	5391	6656	228	71	66	7274	1714	605	232
Setaria italica	39295	2436	785	6748	599	5692	6619	198	62	64	6784	1645	542	225
Sorghum bicolor	33970	2056	625	5211	549	4681	5876	170	51	70	5366	1348	297	242
Glycine soja	32339	1909	615	5436	400	4027	5075	183	40	54	5785	1400	308	127
Glycine glabertima	32094	2151	674	5132	536	4689	4899	187	55	53	5625	1375	518	216
Brachypodium distachyon	30180	2204	527	4415	594	4339	4502	216	59	52	5703	1552	287	208
Hordeum vulgare	21743	1554	538	3397	443	3722	310.	111	37	43	3357	1086	192	186
Dicots														
Glycine max	74114	4369	1004	7759	1256	8955	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	385	106	95	12300	2447	589	148
Arabidopsis lyrata	53847	3666	782	5470	1351	6225	9855	494	464	279	14056	2516	546	251
Populus trichocarpa	45325	2395	512	4750	509	4511	9349	374	138	78	10020	2100	540	179
Solanum lycopersicum	36541	2209	552	3768	525	3817	6816	274	91	76	8264	1808	375	134
Arabidopsis lyrata	32797	2447	383	3404	527	4159	5918	216	90	42	7668	1695	345	192
Picea sitchensis	31471	1848	359	4133	420	3948	5911	217	74	51	6374	1424	322	110
Nelumbo nucifera	26649	1313	448	3696	585	3511	4446	196	59	40	5350	1352	261	98
Lotus japonicus	8674	555	90	1041	172	1071	1701	64	29	17	1585	269	129	47
Mosses														
Physcomitrella patens	34939	761	413	4641	280	3519	9058	287	79	66	7542	1080	218	35
Selaginella moellendorffii	33284	1749	600	4718	317	2739	7694	297	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	66935	25545	173075	47635	237682	252258	9931	2755	2069	247354	54497	13114	5102

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

表 3 绿藻,单子叶植物和双子叶植物中亚细胞蛋白质组分布的比较

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	10271	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 ($t < 0.001$)

为选择。数据库特性的概述如图 1。人为选择的分泌蛋白由从 UniProtKB / Swiss-Prot 检索亚细胞位置标记为“审查”以及人为策划的蛋白质组成。来自内部治疗和团体的选择的蛋白质由其亚细胞位置注释和相关文献的实验依据支持。注释页面还包含主要蛋白质序列(图 1)。

EST 数据注释包含主要 EST 序列，使用 OrfPredictor 预测的蛋白质肽序列(Min et al.,

2005a)，基于 BLASTX 的功能注释，使用目标标识符预测开放阅读框的完整性(Min et al., 2005b)，使用工具生成的亚细胞位置预测相关信息基于预测的蛋白质序列。由于 EST 数据可能包含在测序和装配中引入的错误，需要小心使用数据。数据库中提供的 EST 信息将有助于数据挖掘和实验设计以进一步检查所编码蛋白质的基因功能和亚细胞位置。

2.2 数据总结

PlantSecKB 总共包含 1 415 921 个蛋白质序列，包括来自 UniProt / Swiss-Prot 数据集(选择和审查)的 33 643 个，来自 UniProt-TrEMBL(未审查)的含有 26 685 个从新的圣莲基因序列获得的额外蛋白的 1 355 593 个，(Ming et al., 2013; Lum et al., 2013)。对于具有超过 7 000 个蛋白的物种，亚细胞蛋白质组学的主要类别总结在表 1 中，卷曲的分泌蛋白，OR 蛋白和溶酶体蛋白未列于表 1 中。在拟南芥中只预测到 7 种溶酶体蛋白，其他物种中没有预测到溶酶体蛋白。共有 2 774 个卷曲分泌蛋白，其主要获自拟南芥和油菜亚种粳稻，分别具有 1 247 和 559 个。应当注意的是一个物种的总蛋白质条目是在 UniProtKB 中收集的数目，因为在一些蛋白质条目中存在一些冗余或重复，其可以大于完整或参照基因组。例如，苜蓿亚种在 PlantS 中有 99 984 个条目，在其完整的蛋白质组中只有 63 544 个条目，and 拟南芥在 PlantS 中有 53 847 个条目，在 UniProtKB 完整蛋白质组中只有 31 908 个条目(http://www.uniprot.org/taxonomy/complete-proteomes)。观察到的总体趋势是具有相对小的蛋白质组的植物具有相对较小数量和相对较低比例的分泌的蛋白，例如在单细胞绿藻中。例如，肠球菌属具有少于 100 种预测的分泌蛋白(1.2％)，苔藓（Physcomitrella patens）预测到有 781 种分泌蛋白(2.9％) (表 2)。基于我们的预测估计，单子叶植物和双子叶植物的平均蛋白质组占蛋白质组的 4.0％~7.5％。在这项报告的蛋白质组百分比略低于我们以前报告，这是因为我们以前的研究使用了 SignalP 3.0，而本研究使用了 SignalP 4.0，它具有更高的特异性(Lum et al., 2013; Petersen et al., 2011)。

预测的 9 种列于表 2 中的藻类、单子叶植物和双子叶植物的亚细胞蛋白质组学的平均蛋白质组大小和分布总结在表 3 中。莲花粳稻，一种双子叶植物，由于其蛋白质组的不完全性，是唯一不用于该分析的物种。绿藻中预测的平均蛋白质组要小得多，所以每个亚细胞蛋白质组仅由较少数量的蛋白质组成(表 3)。比较单子叶植物和双子叶植物中分泌蛋白、叶绿体膜蛋白、液泡蛋白和质膜蛋白的分布百分比没有显着差异，但是，预测为线粒体和叶绿体膜蛋白的，单子叶植物具有显著更高比例(包括膜和非膜)；双子叶植物具有显著更多的蛋白质预测为细胞质和核蛋白(表 3)，这些观察到的单子叶植物和双子叶植物之间亚细胞蛋白质组分布的差异是由计算工具还是生物本身或进化意义引起的需要进一步研究。

3 分泌蛋白组的比较分析

植物分泌蛋白或其他亚蛋白质组的完全比较进化分析不在本研究的范围。但是，由于完全分泌蛋白组或其他亚蛋白质组序列可以直接从我们的数据库下载，它将有助于进一步比较研究不同物种中的这些亚蛋白质组。举个例子，我们对包括三种单子叶植物的代表(短柄苋、粳稻、玉蜀黍)、三种双子叶植物(拟南芥、毛果杨、番茄)和两种苔藓(小立碗藓、卷柏)的一组植物进行了分泌蛋白组的比较分析(表 4; 表 5)。我们使用 BLAST 中的 blastclust 工具，在比对中截断 95％的同一性以去除或减少冗余，对非冗余或较少冗余的分泌物进行比较。为了提供植物中分泌蛋白的功能的概述，我们对 8 个选定的植物物种进行了代表性分泌蛋白基因本体(GO)的分析。分泌蛋白组被用来查找具有 1e-10 的截断 E 值的 BLASTP Swiss-Prot 数据集。基因本体(GO)信息从 UniProt ID 映射数据检索(http://www.uniprot.org/downloads)并使用 GO SlimViewer 和植物特异性基因本体(GO)术语进行分析(McCarthy et al., 2006)。所选物种分泌蛋白的 GO 生物过程和分子功能分类的比较总结在表 4 中。植物分泌的蛋白参与超过 40 种不同的生物过程，包括代谢和分解代谢过程、对生物或非生物刺激的应激反应、碳水化合物、脂质和蛋白质代谢过程、多细胞生物体发育等。分子功能分类显示植物分泌蛋白组由大量水解酶(~30％)和转移酶(7％~9％)组成，并且大部分具有各种结合活性(~40％)或催化活性(12％~15％)。应当注意的是，因为许多分泌的蛋白质没有被分类在 GO 中，GO 分类仅是每个类别的分布的估计。

使用 rpsBLAST 在保守结构域数据库(CDD)中

表 4 不同植物物种中分泌蛋白的基因本体分类

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)
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*

搜索 Pfam 进一步分析分泌蛋白的功能(Marchler-Bauer et al., 2009)。Pfam 分析结果中具有 20 个或更多个成员的物种总结在表 5 中。Pfams 的完整列表可以在补充表 1 中找到。搜索 Pfam 的分泌蛋白分子功能的详细分析揭示了不同物种之间的蛋白质家族的差异，包括给定的 Pfam 和物种特异性 Pfam 中的成员数目的变化(表 5)。值得注意的是，水稻中分泌过氧化物酶蛋白的数量是拟南芥的两倍(表 5)。植物过氧化物酶具有多种组织特异性功能，例如从叶绿体和细胞溶质中除去过氧化氢，毒性化合物的氧化，细胞壁的生物合成和针对伤害的防御反应(Sottomayor and Barceló, 2004)。糖基水解酶被认为在修饰植物细胞壁结构和新的生物能源和原料的开发方面有很重要的价值。(Lopez-Casado et al., 2008)。水稻分泌蛋白组由 31 个成员的 Glyco-hydro-18(GH18)和 26 个 GH32N 组成，而在拟南芥分泌组织中仅检测到两个 GH18 和 6 个 GH32N。我们还观察到一些 Pfams 在水稻中有比在其他物种中更多的成员，这些 Pfams 包括 dirigen-like 蛋白，多铜氧化酶，花粉过敏原，细胞

表 5 代表性植物物种的分泌物中蛋白质家族的比较

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
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

色素 P450 等(表 5)。应当注意的是，这些预测的分泌的细胞色素 P450 蛋白很可能是假阳性，因为到目前没有关于在植物中存在分泌的细胞色素 P450 蛋白的有实验依据的报道。Wen 等(2007)曾报道了豌豆根盖分泌蛋白组中存在细胞色素 P450，但是它的存在可能表示在细胞分离过程中发生了泄漏。由于基因组比较小，苔藓物种通常具有较少的分泌蛋白和给定 Pfam 中的较少的成员数，但是我们发现石松门模式生物卷柏有 D-甘露糖结合凝集素家庭的 20 个成员，而其他植物物种在这个 Pfam 中除了毛茛包含 30 个成员，都只有少于 10 个的成员。我们还观察到了物种特异性分泌蛋白，例如玉米具有 30 个玉米醇溶蛋白种子储存蛋白成员；小立碗藓(subsp. patens)具有 61 个具有未知功能的蛋白质成员(DUF4100)。

4 讨论

为了给植物研究界提供资源我们构建了 PlantSecKB。由 UniProtKB 或我们策划的给定蛋白质的亚细胞位置被认为是首次分配的亚细胞位置，这些分配基于具有实验证据的可追溯文献，因此相当可靠。但是，基于计算预测分配的亚细胞位置将取决于所使用的工具的精度。我们已经评估了在本研究中使用的方法的预测准确性，并将其与其他方法的准确性进行了比较(表 1) (Min, 2010; Meinken and Min, 2012)。我们认为分泌蛋白的预测是相对可靠的，但是仍然存在假阳性和假阴性，例如，许多被预测为分泌蛋白的 P450 酶很可能是假阳性。

我们还预测了其他亚细胞位置，包括基于 TargetP 和 WoLF PSORT 的预测的线粒体，叶绿体，液泡，细胞核等。我们对这些亚细胞位置的预测准确度的评估揭示了我们使用的工具的准确性，尽管它们是可用工具中最好的，但因为这些亚细胞位置的预测灵敏度相对较低，仍然不能令人满意(表 1) (Meinken and Min, 2013)。除线粒体和胞质蛋白外，对于叶绿体，ER、高尔基体、细胞核、质膜、液泡和细胞骨架的亚细胞位置特异性是可接受的(>89%)。因此，在那些亚细胞位置预测的蛋白质是相对可靠的，虽然它们仍然需要通过实验验证。最近，几种新工具相继被开发了，包括 Cell-PLoc 服务器(Chou and Shen, 2008)、MultiLoc2 (Blum et al., 2009)以及其他(Meinken and Min, 2012)。这些工具及其相关出版物可以在我们的网站上找到(http://proteomics.ysu.edu/tools/subcell.html)(Meinken and Min, 2012)。由于其中一些工具不能独立使用，比如 Cell-PLoc 服务器，一些可以独立使用的工具则太慢，无法处理大型数据集，比如 MultiLoc2，我们无法使用它们进行数据处理。但是，我们建议用户使用这些工具获得感兴趣的蛋白质的第二预测，因为我们的经验显示使用多个工具可以改善预测特异性。基于最近对植物中分泌物的大规模研究，观察到非经典的，即无铅分泌蛋白(LSP)被占总的鉴定的分泌蛋白组的 50％以上，推测存在独立于经典 ER-高尔基分泌途径的新型分泌机制(Agrawal et al., 2010 for review; Jung et al., 2008; Cheng and Williamson, 2010; Ding et al., 2012)。哺乳动物和细菌 LSPs 已被收集并用预测这些蛋白质的预测软件 SecretomeP 预测(http://www.cbs.dtu.dk/services/SecretomeP/)(Bendtsen et al., 2004a)。因为该工具未处理过植物特异性数据，无法评估预测植物 LSP 的准确性，我们没有在数据处理中使用这个工具。

PlantS eKB 努力成为植物研究人员搜索植物蛋白，特别是分泌蛋白亚细胞位置的门户， EST 子数据库有望促进表达数据的分泌蛋白的 EST 数据挖掘，这对于未完全测序或仅具有有限数目的 cDNA 序列的植物物种特别有用。从收集和整理的具有实验依据的植物分泌蛋白，特别是 LSP 文献来看，植物研究界仍然需要不断努力。我们实施了一个方便人为选择具有实验依据植物蛋白亚细胞位置可通过 PlantSecKB 访问的管理工具，FunSecKB 中描述的实用程序和我们最近实施的真菌分泌蛋白质组知识库(FunSecKB) (Lum and Min, 2011b)提供预期的搜索、下载和管理系统，这将有助于植物研究界进一步了解分泌蛋白质组生物学。它还可以用于探索植物和真菌分泌蛋白的各种潜在作用及其联系、植物病原体控制和抗逆性品种的培育(Kim et al., 2009)。

作者贡献

GL 和 JM 负责数据库，JO 和 SF 进行分泌蛋白的人为选择，XJM 设计和构思了整个实验及设数据处理的流程。XJM，JM 和 GL 负责数据的分析和文章的撰写。所有作者阅读并同意最终的文本。

致谢

本研究由俄亥俄植物生物技术联盟 [授权

2011-001] (俄亥俄州立大学，俄亥俄州农业研究和发展中心)和斯敦州立大学(YSU)研究理事会[授权 2010-2011 和 12-11]共同资助。本研究还得到了扬斯敦州立大学(YSU)研究教授以及科学、技术、工程学院的支持，数学院院长将研究的时间分配给 XJM。JM 由扬斯敦州立大学(YSU)应用化学生物学中心的研究生助手提供支持。

参考文献

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

Alexandersson E., Ali A., Resjö S., and Andreasson E., 2013, Plant secretome proteomics, Front. Plant Sci., 4: 9

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

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

Blum T., Briesemeister S., and Kohlbacher O., 2009, MultiLoc2: integrating phylogeny and Gene Ontology terms improves subcellular protein localization prediction, BMC Bioinformatics, 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

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

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

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

Cheng F.Y., and Williamson J.D., 2010, Is there leaderless protein secretion in plants? Plant Signal Behav., 5: 129-131

Cho W.K., and Kim J.Y., 2009, Integrated analyses of the rice secretome, Plant Signal Behav., 4: 345-347

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

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

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

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

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

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

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

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

Foresti O., and Denecke J., 2008, Intermediate organelles of the plant secretory pathway: identity and function, Traffic, 9: 1599-1612

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

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

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

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

Jones R.L., and Robinson D.G., 1989, Protein Secretion in Plants, Tansley Review No. 17, New Phytologist, 111: 567-597

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Sottomayor M., and Barceló A.R., 2004, Plant peroxidases and phytochemistry – foreword, Phytochemistry Rev., 3: 1-2

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

von Heijne G., 1990, The signal peptide, J. Membr. Biol., 115: 195-201

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

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