计算分子生物学(网络版), 2015 年, 第 4 卷, No.10, 1-11
Jisuan Fenzi Shengwuxue (Online), 2015, Vol.4, No.10, 1-11
http://cmb.biopublisher.cn

研究论文
Research Article

FunSecKB2: 真菌蛋白亚细胞定位的知识库

John Meinken[1,4] David K.Asch[2,4] Kofi A.Neizer-Ashun[3] Guang-Hwa Chang[3] Chester R.Cooper JR[2,4] Xiang Jia Min [2,4*]

1 计算机科学与信息系统系, 扬斯敦州立大学, 杨斯顿, 美国

2 生物科学, 扬斯敦州立大学, 扬斯敦, 美国

3 数学系, 扬斯敦州立大学, 杨斯顿, 美国

4 应用化学与生物学中心, 扬斯敦州立大学, 扬斯敦, 美国

通讯作者: xmin@ysu.edu

计算分子生物学, 2015 年, 第 4 卷, 第 10 篇 doi: 10.5376/cmb.cn.2015.04.0010

建议最佳引用格式:

Meinken et al., 2014, FunSecKB2: a fungal protein subcellular location knowledgebase, Computational Molecular Biology, Vol.4, No.6, 1-17 (doi: 10.5376/cmb.2014.04.0007)

摘 要 FunSecKB2 是真菌分泌和亚细胞蛋白质组的改进和更新版本，即蛋白质的亚细胞定位的知识库。真菌蛋白序列数据取自 UniProtKB 数据库，由近 200 万个与 167 种完整的蛋白质组相关的蛋白质序列组成。基于精确的信息和其中计算的工具来进行蛋白质的亚细胞定位。用于亚细胞定位预测的工具包括 SignaIP、WoLF PSORT、Phobius、TargetP、TMHMM、FragAnchor 和 PS-Scan。分泌蛋白即分泌蛋白组，连同其他 15 个亚细胞蛋白质组进行预测。用户可以通过搜索几种不同的标识符在该数据库中搜索到所需数据，可以是基因名称或关键字(词)。任何一个物种的亚细胞蛋白质组可以被搜索或下载，而 BLAST 搜索整个真菌蛋白数据或分泌蛋白数据是可用的。基于实验证据的亚细胞位置的群落注释也是被支持。初步分析显示，一种真菌的分泌蛋白组大小是其生活方式的决定因素之一。基因本体论和蛋白质结构域分析显示，真菌分泌蛋白组中包含很多水解酶类、肽酶、氯化还原酶类和裂解酶，这可能在化学废料和生物燃料生产的生物处理中具有潜在的应用。该数据库为真菌群落寻找蛋白质的亚细胞定位信息和亚细胞蛋白质组分析提供了一个重要的且丰富的资源。

关键词 计算预测, 真菌, 分泌蛋白, 分泌, 信号肽, 亚细胞定位, 亚细胞蛋白质组

FunSecKB2: a fungal protein subcellular location knowledgebase

John Meinken [1,4] David K.Asch[2,4] Kofi A.Neizer-Ashun[3] Guang-Hwa Chang[3] Chester R.Cooper JR [2,4] Xiang Jia Min [2,4*]

1 Department of Computer Science and Information Systems, Youngstown State University, USA
2 Department of Biological Sciences, Youngstown State University, Youngstown, USA
3 Department of Mathematics & Statistics, Youngstown State University, Youngstown, USA
4 Center for Applied Chemical Biology, Youngstown State University, Youngstown, USA
Corresponding author, xmin@ysu.edu

Abstract FunSecKB2 is an improved and updated version of the fungal secretome and subcellular proteome, i. e. protein subcellular location, knowledgebase. The fungal protein sequence data were retrieved from UniProtKB, consisting of nearly 2 million entries with 167 species having a complete proteome. The assignments of protein subcellular locations were based on curated information and prediction using seven computational tools. The tools used for subcellular location prediction include SignalP, WoLF PSORT, Phobius, TargetP, TMHMM, FragAnchor, and PS-Scan. Secreted proteins, i.e. secretomes, along with 15 other subcellular proteomes were predicted. The database can be searched by users using several different types of identifiers, gene name or keyword (s). A subcellular proteome from a species can be searched or downloaded. BLAST searching whole fungal protein data or secretomes is available. Community annotation of subcelluar locations based on experimental evidence is also supported. A primary analysis revealed that the secretome size of a fungal species is one of the determining factors to its lifestyle. The Gene Ontology

收稿日期：2014 年 8 月 5 日
接受日期：2014 年 9 月 21 日
发表日期：2014 年 11 月 7 日

and protein domain analysis of fungal secretomes revealed that fungal secretomes contain a large number of hydrolases, peptidases, oxidoreductases, and lysases, which may have potential applications in bio-processing of chemical wastes or biofuel production. The database provides an important and rich resource for the fungal community looking for protein subcellular location information and performing comparative subcellular proteome analysis.

Keywords Computational prediction, Fungi, Secreted protein; Secretome, Signal peptide, Subcellular location, Subcellular proteome

真菌在自然界和我们的日常生活中扮演着重要的角色。在自然界中，真菌作为分解者的生物角色，对碳和养分循环至关重要。在我们的日常生活中，食用菌是众所周知的真菌的例子。酿酒酵母，被称为面包酵母，广泛用于酿酒、烘焙和酿造。一些真菌也被称为生产者的药物，如抗生素。真菌在昆虫、动物、人类和植物中也是重要的病原体。

真菌属于真核生物的四个主要分类之一，真菌细胞包含多个亚细胞区，用于执行不同的亚细胞活动。例如，一个线粒体，这是一个封闭的膜结构，主要是用来提供细胞能量，而核是一个储存遗传材料和控制基因转录的地方。在这个过程中，我们规定在一个物种中，所有蛋白分泌到质膜外的叫分泌蛋白组。这些蛋白质包括细胞壁蛋白、细胞外基质蛋白和分泌的可溶性蛋白，其可以作为一种激素或信号分子或酶。然而，分泌途径中发挥转运功能的蛋白不包括在内，因为该类蛋白与我们对分泌蛋白的定义是不一致的(Tjalsma et al., 2000; Lum and Min, 2011a)。在活体营养真菌中，分泌蛋白作为植物和真菌直间的致病或共生的相互作用的主要负责效应是确定的(Girard et al., 2013)。腐生菌大量分泌的水解酶，如糖苷水解酶家族分解纤维素和木质复合材料(Martinez et al., 2004; Martinez et al., 2009; Murphy et al., 2011)。近年来，随着许多真菌的全基因组测序，利用计算和实验方法，对真菌分泌组的蛋白进行鉴定和分析已成为一个重要的研究课题(Bouws et al., 2008)。例如，在以下已被报道的真菌分泌蛋白组包括 *Aspergillus niger* (Tsang et al., 2009; Braaksma et al., 2010)，*Aspergillus fumigatus* (Powers-Fletcher et al., 2011)，*Candida albicans* (Lee et al., 2003; Ene et al., 2012)，*Doratomyces stemonitis* C8 (Peterson et al., 2011)，*Fusarium graminearum* (Paper et al., 2007; Brown et al., 2012)，*Irpex lacteus* (Salvachúa et al., 2013)，*Magnaporthe oryzae* (Jung et al., 2012)，*Mycosphaerella graminicola* (Morais et al., 2012)，Paracoccidioides (a complex of several phylogenetic species) (Weber et al., 2012)，*Penicillium echinulatum* (Ribeiro et al., 2012)，*Phanerochaete chrysosporium* (Wymelenberg et al., 2005)，*Sclerotinia sclerotiorum* (Yajima and Kav, 2006)，*Trichoderma harzianum* (Do Vale et al., 2012)和 *Ustilago maydis* (Mueller et al., 2008)。

两个真菌特异性分泌蛋白质组数据库，真菌分泌蛋白质组数据库(FSD, http://fsd.snu.ac.kr/)和真菌分泌的知识库(FunSecKB, http://proteomics.ysu.edu/secretomes/fungi.php)已建成，用于搜索真菌分泌的相关信息(Choi et al., 2010; Lum and Min, 2011)。FSD 是基于 9 个不同的程序构造，采用三层递阶辨识规则而建(Choi et al., 2010)。我们开发的 FunSecKB，使用 6 种不同的工具从真菌的参考数据中预测真菌的分泌蛋白(Lum and Min, 2011)。不管怎样，自从 FunSecKB 的发布，可用真菌蛋白数据已大大增加。在这项工作中，我们描述了 FunSecKB2，真菌蛋白亚细胞定位的知识库，也被称为真菌分泌和亚细胞蛋白质组知识库(2 版)，即 FunSecKB 的扩大、更新和改进版。FunSecKB2 由一个精准的协议组成，包括亚细胞信息、分泌蛋白组的预测信息和 15 个亚细胞蛋白组的定位信息。这种改进的真菌蛋白库预计将作为一个中央门户网站，为真菌研究和工业界的用户提供真菌蛋白的亚细胞位置信息，就是那些对利用真菌促进全球生物经济发展感兴趣的人(Lange et al., 2012)。

1 数据采集与数据库的安装

1.1 数据采集

所有真菌蛋白序列从 UniProtKB/Swiss-Prot 数据库和 UniProtKB / TrEMBL 数据库检索到的(2013 年 8 月发布, HTTP://www.uniprot.org/downloads)。对 UniProtKB/Swiss-Prot 数据库主要包含一些非冗余蛋白序列的注释信息，这些信息是从文献的实验结果和经过鉴定的计算分析结果中获取的(UniProt, 2014)。UniProtKB/TrEMBL 数据库包含有蛋白质序列相关的计算生成的注释信息和大量的蛋白功能特性。该数据集包括 1 976 832 个真菌蛋白序列，

其中从 UniProtKB/Swiss-Prot 数据库获取了 30859 条，从 TrEMBL 数据库检索到 1 945 973 条。

1.2 蛋白质亚细胞定位分配的方法

真菌蛋白序列，使用以下程序处理：SignalP (版本 3 和 4, http://www.cbs.dtu.dk/services/SignalP/) (Bendtsen et al., 2004b; Petersen et al., 2011), Phobius (http://phobius.binf.ku.dk/) (Käll et al., 2007), WoLF PSORT (http://wolfpsort.org/) (Horton et al., 2007)和 TargetP (http://www.cbs.dtu.dk/services/targetp/) (Emanuelsson et al., 2007)，进行信号肽和亚细胞定位的预测。这些预测信息顺利通过了预先的评估，得到了真菌分泌研究界广泛应用(Min, 2010)。TMHMM (http://www.cbs.dtu.dk/services/tmhmm)是用来识别具有跨膜结构域蛋白(Krogh et al., 2001)，Scan Prosite (PS-Scan, http://www.expasy.org/tools/scanprosite/)被用来扫描内质网(ER)的靶序列(序列: ps00014) (de Castro et al., 2006; Sigrist et al., 2010)。利用 TMHMM 预测膜蛋白，具有膜结构域的区段并不位于 N-末端(70 个氨基酸)却被视为真正的膜蛋白。利用 SignalP (version3)对蛋白质序列进行预测，发现有一个信号肽使用，使用 FragAnchor 进一步识别为糖基磷脂酰肌醇 (GPI) 锚点 (http://navet.ics.hawaii.edu/~fraganchor/NNHMM/NNHMM.html) (Poisson et al., 2007)。除了 FragAnchor，我们还使用本地 Linux 系统作为独立的工具进行数据的处理。如何通过命令的操作来控制这些分析工具的运行，通常可以在下载的“Readme”的工具包中进行查看(Lum and Min, 2013。

真菌蛋白的亚细胞定位的类别包括：分泌蛋白、线粒体(细胞膜或无膜)、内质网(膜或腔)、细胞溶液(细胞质)、细胞骨架、高尔基体(膜或腔)、细胞核(膜或无膜)、液泡(膜或无膜)、溶酶体、过氧化物酶体、质膜和其他膜蛋白。为了给予蛋白亚细胞定位信息，在使用预测信息之前，我们要考虑 UniProtKB 的注释信息和我们筛选的亚细胞定位信息。对于没有注释亚细胞信息的蛋白质，它们的亚细胞定位主要是基于预测信息。我们最近的计算工具精度评价表明，对真菌分泌蛋白组预测精度最高(92.1%的敏感性和特异性 98.9%)的分析工具组合是 SignalP、WoLF PSORT，结合 Phobius 有利于对信号肽的预测，结合 TMHMM 有利于对膜蛋白的预测，结合 PS-Scan 有利于去除内质网靶蛋白(Min., 2010)。因此，分泌蛋白组分析是受限的，主要是存在人为精选的分泌蛋白，以及利用三个分析工具预测的蛋白 N-端具有信号肽，但没有一个跨膜结构域或内质网导向信号。在这项研究工作中，SignaIP4 代替 SignaIP3 以提高预测精度(Petersen et al., 2011; Melhem et al., 2013)。然而，所产生的 SignaIP3 对于信号肽裂解位点的预测信息仍然要比 SignaIP4 准确(Petersen et al., 2011)。蛋白质的亚细胞定位的详细方法如下所述。

分泌蛋白

分泌蛋白进一步划分为原的分泌蛋白、极有可能会分泌的蛋白、可能会分泌的蛋白，和弱可能分泌蛋白。原始的分泌蛋白包括在 UniProtKB/Swiss-Prot 亚细胞定位中被注释为“分泌”或“细胞外”或“细胞壁”的蛋白。它还包括我们研究人员从最近的文献中，手动收集的分泌蛋白。SignaLP4、Phobius 和 WoLF PSORT 这三个预测分析工具用于分泌信号肽或蛋白质的亚细胞定位预测。极有可能的分泌、可能的分泌和弱可能的分泌蛋白通过这三个预测分析工具能够预测到会分泌或者分泌信号肽。这些蛋白质并没有一个跨膜结构域或是内质网滞留信号肽。

内质网蛋白

利用 WoLF PSORT 和 PS-Scan 对内质网蛋白进行预测分析。预测的蛋白质包括 SignalP4 预测的信号肽，PS-Scan 预测的内质网靶标多肽是内质网管腔蛋白。此外，如果它们包含一个或多个跨膜结构域，它们也被归类为内质网膜蛋白。

GPI 锚定蛋白

信号肽包含的蛋白质被 FragAnchor 预测为含有 GPI 锚点，进一步划分为 GPI 锚定蛋白。蛋白质序列预测到具有信号肽，GPI 锚定蛋白能有与质膜或者变为细胞壁的分泌成分结合。

其他亚细胞位置的蛋白质

其他亚细胞位置，包括线粒体、细胞溶液(细胞质)、细胞骨架、高尔基体、溶酶体、细胞核、过氧化物酶体、质膜和液泡的蛋白通过 WoLF PSORT 来预测。对于预测位于线粒体，高尔基体，细胞核和液泡的蛋白质，如果一个蛋白质包含一个或多个跨膜结构域，它被进一步列为一个膜蛋白在特定的亚细胞定位。

1.3 数据库的安装

数据存储在一个 Linux 服务器 MySQL 关系数据库。用户界面和模块访问数据采用 PHP 模式。从

主用户界面上的链接 http://proteomics.ysu.edu/secretomes/fungi2/index.php.可以访问 BLAST 程序以及提交注释信息。补充表和其他数据可以在 http://proteomics.ysu.edu/publication/data/FunSecKB2/.下载。

2 结果

2.1 蛋白质亚细胞定位预测精度的评价

我们采用了上述基于我们以前的计算工具的预测评价方法(Min, 2010; Meinken and Min, 2012; Melhem et al., 2013)。为了进一步评估我们的方法的预测精度，我们从 UniProtKB/Swiss-Prot 数据库中获得了 14884 个已经完成注释、具有唯一亚细胞定位信息的蛋白。如果蛋白质有多个亚细胞定位或片段化标签，我们就将其排除在外。基于我们之前的公式算法，我们对预测精度的灵敏度、特异性和 MCC 系数进行了测量(Min, 2010)。精度结果如表 1 所示，由于作为阳性蛋白的样本数量太少(＜20)，所以质膜和溶酶体的预测精度并未列入在内。相较于使用单一的工具方法，我们的方法使用多种工具包括 SignalP4、WoLF PSORT 和 Phobius 来预测分泌的信号肽，用 PS-Scan 去除内质网蛋白，TMHMM 用于去除膜蛋白，这些方法能够显著提高预测精度(Min, 2010; Meinken and Min, 2012)。在对某一种蛋白质大小进行预测时，预测的高度可能分泌蛋白将会得到一个相对准确的估计，这一方法具有最高的特异性(> 0.99)，有趣的是，假阴性的数量是接近用于评价的数据集的假阳性的数量。包括预测可能为分泌蛋白组其 MCC 值仅略有下降，的只有少数蛋白条带属于这一类。然而，对于弱可能分泌型蛋白的预测分析需要谨慎对待，因为假阳性的数量远远超过的假阴性数量的减少(表 1)。

表 1 真菌蛋白亚细胞定位预测精度的评价

Table 1 Evaluation of prediction accuracies of fungal protein subcellular locations

Subcellular location	True positive	False positive	True negative	False Negative	Sn	Sp	MCC
HLS	1364	130	13269	121	0.919	0.990	0.906
HLS+LS	1401	188	13211	84	0.943	0.986	0.902
HLS+LS+WLS	1412	337	13062	73	0.951	0.975	0.862
Mitochondria	1595	887	12015	387	0.805	0.931	0.671
ER	19	11	13873	981	0.019	0.999	0.102
Golgi apparatus	5	2	14527	350	0.014	1.000	0.098
Nucleus	4483	2771	6823	807	0.847	0.711	0.535
Vacuole	0	0	14389	495	0.000	1.000	
Peroxisome	9	15	14722	138	0.061	0.999	0.148
Cytoplasm	1293	762	10611	2218	0.368	0.933	0.371
Cytoskeleton	87	234	14055	508	0.146	0.984	0.175

注: HLS: 高度可能的分泌; LS: 可能分泌; WLS: 弱可能分泌; ER: 内质网; Sn: 敏感性; SP: 特异性; MCC:马休斯相关系数

Note: HLS: highly likely secreted; LS: likely secreted; WLS: weakly likely secreted; ER: Endoplasmic reticulum; Sn: sensitivity; Sp:specificity; MCC: Matthews correlation coefficient

我们还比较了线粒体蛋白由 WoLF PSORT 和 TargetP 预测的准确性。我们发现，其 Wolf PSORT 预测的 MCC 值为 0.67，TargetP 预测的 MCC 值为 0.56。我们也发现了使用所有的预测分析工具增加线粒体蛋白预测的特异性，当同时使用时，MCC 值从单一使用 WoLF PSORT 时的 0.93 增加到> 0.98。然而，由于预测的敏感度降低，使用的所有的分析工具并不能提高 MCC 的值。因此，我们选择了 WoLF PSORT 对线粒体蛋白进行定位。然而，用户应该意识到，如果 WoLF PSORT 和 TargetP 预测蛋白都是一种线粒体蛋白，那么这个预测结果要比单一的预测结果可靠得多。

其他亚细胞位置的预测精度存在很大的差异。核蛋白的预测的敏感性值为 0.85 特异性值为 0.71，MCC 值为 0.53。其他的亚细胞位置包括内质网、高尔基体、液泡、细胞质和过氧化物酶体等的预测

精度在 MCC 值上很低(＜0.4) (表 1)。然而，应该指出的是，低的精度所造成的灵敏度也是非常低的，事实上，其特异性却是比较高的(＞0.98)。因此，会有一部分的位于这些亚细胞位置的蛋白是无法检测到的。然而，如果一个蛋白质被预测为位于这样的位置，那么这个预测结果是最有可能是正确的。但是，预测这些真菌蛋白的亚细胞位置的精度仍需要提高。

2.2 不同物种的亚细胞蛋白质组分布概况

该数据库包含预测的蛋白质的亚细胞位置信息，来自于 16554 个真菌物种或品种(株)，其中有 189 个每个至少有 1000 条蛋白质序列。物种名称，其中一些可能有一个以上名称，可以在用户界面上看到，方便特定物种的搜索或下载。具有大于 1000 条蛋白质序列的物种，也可以通过由用户提供的物种名称搜索到。不同真菌中物种的亚细胞蛋白质组的分布情况总结在表 2 和附表 1。表 2 包括以下的亚细胞位置：分泌蛋白(4 类)、线粒体膜和线粒体无膜、细胞溶液(细胞质)、细胞骨架、核膜及核无膜、质膜与糖基磷脂酰肌醇(GPI)锚定蛋白。分泌蛋白的种类包括以下类别：原始的分泌、极有可能会分泌、分泌和弱分泌蛋白。其他亚细胞蛋白质的位置信息(包括内质网的膜或腔)、高尔基体(膜或腔)、溶酶体、过氧化物酶体、液泡(膜或无膜)，其他的膜，和其他的位置信息，在附表 1 中。

基因组大小是易变的，因此，在不同的真菌物种中，蛋白质组的大小也是相当大的。然而，应该指出的是，在数据库中，如表 2 所示，一个给定的物种的总蛋白不一定是蛋白质组的大小，而是一个从不同物种得到的蛋白质的集合。例如，酿酒酵母的蛋白质组的大小， UniProtKB 数据库中只包括 6621 的蛋白质，但是在我们的数据库中，有 79093 个蛋白质归类到酿酒酵母的名下，明显说明这些蛋白是来自于多个菌株。基于汇总数据为每个子囊菌门、担子菌门和微孢子虫评估真菌蛋白的亚细胞分布。有趣的是，我们发现核是最大亚细胞定位位置：子囊菌门有 39.2%蛋白质定位在核内，担子菌门有 39.2%的蛋白质，在微孢子虫有 57.4%的蛋白质定位在核内。线粒体蛋白是另一个大的定位区域：子囊菌门有 19.5%的蛋白，担子菌门有 21.1%的蛋白，微孢子虫有 16.7%的蛋白，分别位于线粒体。约 18~21%的蛋白质位于细胞质或细胞质中。在子囊菌门，分泌蛋白组的比例从 0.3%到 10.5%不等，平均含量为 4.6%；担子菌门从 1.9%到 7.4%不等，平均含量为 4.4%；微孢子虫从 0.5%到 1.7%不等，平均含量为 0.9%。然而，这里的分泌蛋白组也是有限的，包括原始的分泌蛋白质和极有可能的分泌蛋白，因此一个数字代表一种蛋白组。包括其他预测的蛋白质，分泌性蛋白和弱分泌性蛋白，分泌蛋白组的大小会显著增大。预测弱可能分泌分泌蛋白的分泌可能规模肯定会明显增加，但也会有假阳性和非分泌型蛋白。

2.3 不同真菌的生活史与分泌蛋白组的关系

类似于我们以前 FunSecKB 的分析工作(Lum and Min, 2011)，在一个物种中分泌蛋白组的大小(y)与蛋白质组的大小高度相关(x) (r= 0.87)，回归方程为 y= 0.081×271 (图 1)。然而，物种不同的生活方式在分泌蛋白组与分泌蛋白的比例大小存在差异。Lowe 和 Howlett (2012)研究真菌的生活方式和分泌蛋白组的大小，发现具有双相生活方式的真菌菌有很大比例的分泌蛋白，相较于腐生菌或植物真菌，动物病原体只有少部分的基因被预测具有分泌蛋白。在 Lowe 和 Howlett 的工作(2012)中，该组仅用 SignalP 进行分泌蛋白组的预测，因而，其大小可能超过估计。利用我们在这项工作中收集到的数据，我们研究了真菌的生活方式和其分泌的尺寸之间的关系(图 1 和附表 2)。

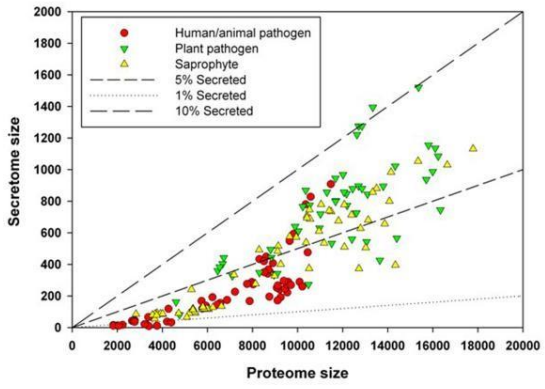

图 1 蛋白质大小和真菌物种具有不同的生活方式之间的关系

Figure 1 Relationship between proteome size and secretome size in fungal species having different lifestyles

在数据库中的每个物种的数据包含冗余或重复蛋白的条目，我们只是用 UniProt 数据库中的蛋白质或完整的蛋白质组数据集做为参考(http://www.uniprot.org/taxonomy/complete-proteome

S)。我们收集的物种具有完整的蛋白质组，具有植物或动物病原体、人类的病原体或腐生植物的一种生活方式。其中的一些可能被分为多个类别，这些条目被注释(见补充表 2)。根据 Lowe 和 Howlett 的报道(2012)，人类和动物的病原体，包括病原微生物和一些杀线虫寄生真菌有一个相对较小的蛋白质组，其大小大多都小于 12000 条蛋白质序列，其中一些被称为微孢子虫的寄生虫的基因组编码共有 2000~4000 的蛋白质，不超过 1%的会被分泌(图 1)。人类/动物的病原体，分泌的蛋白质的比例变化从 0.3 到 7.9%，平均为 2.8%。另一方面，植物病原菌和腐生菌有更多的大小从 4000 到 18000 个蛋白质和较高程度的分泌蛋白，在腐生菌中从 1.3%到 7.1%，平均含量为 4.2%，植物病原菌从 1.7%到 10.5%，平均含量为 6.3%。显然，这些结果表明，分泌的大小是控制真菌的生活方式的重要决定因素之一。然而，具有相似大小的分泌蛋白组的物种却可能有不同的生活方式，在每个物种中，分泌蛋白组的组成可能扮演着重要的作用，决定了物种的生活方式。

表 3 真菌分泌蛋白的 GO 分类

Table 3 Gene Ontology (GO) classification of fungal secreted proteins

GO ID	Count	%	GO description
Molecular function			
GO:0016798	16132	30.9	hydrolase activity, acting on glycosyl bonds
GO:0043167	11011	21.1	ion binding
GO:0008233	8182	15.7	peptidase activity
GO:0016491	7305	14.0	oxidoreductase activity
GO:0016829	1710	3.3	lyase activity
GO:0016791	1439	2.8	phosphatase activity
GO:0016810	1242	2.4	hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds
GO:0016853	1010	1.9	isomerase activity
Others	4136	7.9	including 32 other GO categories
Total	52167		
Biological process			
GO:0009056	21356	24.6	catabolic process
GO:0005975	19039	22.0	carbohydrate metabolic process
GO:0071554	5584	6.4	cell wall organization or biogenesis
GO:0009058	3612	4.2	biosynthetic process
GO:0006629	3463	4.0	lipid metabolic process
GO:0006950	3405	3.9	response to stress
GO:0044281	3356	3.9	small molecule metabolic process
GO:0034641	3076	3.5	cellular nitrogen compound metabolic process
Others	23845	27.5	including 60 other GO categories
Total	86736		

2.4 真菌分泌蛋白的功能分析

为了提供所有的真菌分泌蛋白的功能的概述，我们进行了 GO 分析。分泌蛋白组包括原始和预测很有可能分泌的蛋白，仅仅只是为了在 UniProtKB/Swiss-Prot 数据库执行 BLASTP 搜索，E 值设定为 1e−10。GO 注释的信息从 Uniprot ID 映射数据检索(http://www.uniprot. org/downloads)中获取，再利用 GO SlimViewer 分析(McCarthy et al., 2006)。分泌组蛋白 GO 分析产生的生物学过程和分子功能归类于表 3。分子功能分类显示，真菌分泌蛋白是

由大量的水解酶(约 33.7%)、具有离子结合功能的蛋白(约 21.1%)、肽酶(15.7%)、氧化还原酶类(14%)和其他的一些酶。真菌分泌的蛋白质参与了超过 60 个不同的生物过程。主要的生物过程，包括代谢过程(24.6%)、碳水化合物的合成(22%)或脂质代谢过程(4%)，以及细胞壁组织的合成(6.4%)、应激反应、小分子和氮代谢等过程，应该指出的是，GO 分类只是对没有 GO 功能注释信息的约 54%的预测分泌蛋白进行了功能分布的预测

表 4 真菌中高度编码的分泌蛋白家族

Table 4 Highly encoded secreted protein families in fungi

Pfam ID	Members	%[a]	Pfam	Function
pfam00026	1473	3.4	Asp	Eukaryotic aspartyl protease
pfam00135	1419	3.2	COesterase	Carboxylesterase family
pfam01565	1395	3.2	FAD_binding_4	FAD binding domain
pfam00082	1279	2.9	Peptidase_S8	Subtilase family
pfam03443	1150	2.6	Glyco_hydro_61	Glycosyl hydrolase family 61
pfam00295	924	2.1	Glyco_hydro_28	Glycosyl hydrolases family 28
pfam00704	924	2.1	Glyco_hydro_18	Glycosyl hydrolases family 18
pfam05199	873	2.0	GMC_oxred_C	GMC oxidoreductase
pfam00450	845	1.9	Peptidase_S10	Serine carboxypeptidase
pfam00933	809	1.8	Glyco_hydro_3	Glycosyl hydrolase family 3 N terminal
pfam04389	695	1.6	Peptidase_M28	Peptidase family M28
pfam07732	651	1.5	Cu-oxidase_3	Multicopper oxidase
pfam00264	631	1.4	Tyrosinase	Common central domain of tyrosinase
pfam04616	591	1.3	Glyco_hydro_43	Glycosyl hydrolases family 43
pfam01083	569	1.3	Cutinase	Cutinase
pfam09286	519	1.2	Pro-kuma_activ	Pro-kumamolisin
pfam01522	486	1.1	Polysacc_deac_1	Polysaccharide deacetylase
pfam00150	454	1.0	Cellulase	Cellulase (glycosyl hydrolase family 5)
pfam09362	450	1.0	DUF1996	Domain of unknown function (DUF1996)
pfam00328	417	0.9	His_Phos_2	Histidine phosphatase superfamily (branch
pfam00840	410	0.9	Glyco_hydro_7	Glycosyl hydrolase family 7
pfam00188	400	0.9	CAP	Cysteine-rich secretory protein family
pfam01764	397	0.9	Lipase_3	Lipase (class 3)
fam00544	393	0.9	Pec_lyase_C	Pectate lyase
pfam00331	381	0.9	Glyco_hydro_10	Glycosyl hydrolase family 10
pfam00457	377	0.9	Glyco_hydro_11	Glycosyl hydrolases family 11
pfam00246	348	0.8	Peptidase_M14	Zinc carboxypeptidase
pfam12708	337	0.8	Pectate_lyase_3	Pectate lyase superfamily protein
pfam07519	331	0.8	Tannase	Tannase and feruloyl esterase
pfam00722	325	0.7	Glyco_hydro_16	Glycosyl hydrolases family 16
pfam00394	303	0.7	Cu-oxidase	Multicopper oxidase
pfam13668	301	0.7	Ferritin_2	Ferritin-like domain

Note: a, 百分比(%)是基于总的 43853 种蛋白质具有 Pfam 匹配算来计算，完整的列表可以下载
Note: a The percentage (%) was calculated based on a total of 43853 proteins having a Pfam match. The complete list can be downloaded (see text for details)

我们进一步对预测的分泌真菌蛋白进行功能预测，使用 rpSBLAST 工具对 Pfam 数据库进行搜索，E 值设定为 1e-10。在 93430 个预测的分泌蛋白中，43953 个蛋白序列在 Pfam 分析中能够匹配，共有 880 个蛋白质家族进行了检测。真菌中 33 个高度编码的分泌蛋白家族进行了汇总，如表 4 所示，其完整的列表可以下载(http://proteomics.ysu.edu/publicaiton/ data/)。真菌中 10 个高编码分泌蛋白家族有天冬氨酰蛋白酶家族、羧酸酯酶家族、FAD 结合结构域、枯草杆菌蛋白酶家族、糖基水解酶家族 61、糖基水解酶家族 28、糖基水解酶家族 18、GMC 氧化还原酶、丝氨酸羧基肽酶和糖基水解酶 3 家族。这些蛋白酶，例如在这里确定的如天冬氨酸蛋白酶、枯草杆菌蛋白酶、肽酶和其他蛋白家族对各种生长介质或环境中的基质材料存在的蛋白的协同降解可能是必须的(Druzhinina et al. 2012; Girard et al. 2013)。GO 分析和功能结构域的分析结果是一致的，显示这些蛋白主要参与降解复杂的生物分子，包括碳水化合物、蛋白质、脂类和其他分子。

3 讨论

我们建立了一个真菌蛋白亚细胞定位数据库并命名为真菌分泌蛋白质组和亚细胞蛋白知识库(FunSecKB2)。与 FunSecKB 相比，蛋白的总条数从 478073 增加到了 1976832，而真菌种类的数量(包括不同种和不同菌株)从 52 增加到 210。而亚细胞定位也从蛋白质组扩增到包括线粒体、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。此外，对于分泌蛋白质组，我们进一步将其细分，极有可能分为极有可能的分泌蛋白、可能的分泌蛋白、弱可能的分泌蛋白。这种细化的分类和其它亚细胞定位将会大大提高在不同物种中进行的亚细胞蛋白质组的比较分析。然而，当在 UniProtKB 获得蛋白序列后，一些重复的条目也会获得，因此对于一个给定物种的蛋白质组分析，我们需要用到非冗余参考和蛋白质组数据集，这也可以在 UniProt 上下载到(http://www.uniprot.org/taxonomy/complete-proteomes)。还应注意的是，对于列表中一个给定的物种，如果没有特别指定是哪个菌株或者哪个亚种，那么得到的条目将会包含所有蛋白条。

我们还提供了能搜索所有真菌蛋白或者预测的真菌分泌蛋白数据的 BLAST 工具，包括能搜索到他们的序列。这有助于鉴别同源蛋白及其可能的亚细胞定位。另外，任何匿名的蛋白序列使用者都能够利用这个预测蛋白亚细胞定位的工具。其它预测蛋白质组的亚细胞定位的有效工具已经由 Meinken and Min (2012) 和 Caccia 等人(2013)总结过了。近来，Cortázar 等人(2014)应用了一个 web 服务器 SECRETOOL，它集合了多种预测真菌分泌蛋白质组的工具。其中有些工具跟我们使用的一样，我们期望这个服务器能产生相对可靠的预测结果，它也在近来的蛋白质组预测中起到很好的作用(Cortázar et al., 2013; Lum and Min, 2011)。此外，另一个可靠的数据库，真菌分泌蛋白质组数据库(FSD)，是用一种不太一样的工具建立而成的，它也可能提供其它一些真菌蛋白的亚细胞定位的信息。

真菌有一套适应周围环境的分泌系统，而环境约束导致的选择压力使得物种进化出各种各样的复杂的蛋白质组(Girad et al., 2013; Alfaro et al., 2014)。由于生活方式的不同，腐生性生物的蛋白质组中主要有水解酶。活体生物既有降解水解酶，；菌根生物既有水解酶、又有生物相容性、转换效应蛋白，死体营养型生物有水解酶和致死蛋白(Girad et al., 2013, Alfaro et al., 2014)。分泌蛋白质组基本包含两类蛋白：大部分是以多糖水解酶为代表的蛋白，例如作用于糖苷键的水解酶。还有一小部分是包括蛋白酶、脂肪酶和氧化还原酶等(表 3)。在这项工作中，对蛋白的鉴定仅限于经典的分泌蛋白，如含有蛋白的信号肽和那些包含经典蛋白和无铅分泌蛋白。SecretomeP 是一个被用来预测细菌和哺乳动物的无铅分泌蛋白的工具(http://www.cbs.dtu.dk/services/SecretomeP/)(Bendtsen et al., 2004a)。因为这个工具没有被用过预测真菌的数据，预测的精确性无法评估，因此我们没有用这个工具来处理数据。我们想请真菌研究团队提交下真菌蛋白的亚细胞定位数据，特别是无铅分泌蛋白，而且是有文献中的实验数据支持的数

据。对一个物种蛋白质组的计算机预测为实验验证和鉴定不同环境或者培养条件下的分泌蛋白提供了第一步的工作基础(Alfaro et al., 2014)。随着我国出版的植物分泌和亚细胞蛋白数据库 (PlantSecKB) (Lum et al., 2014)，我们期望 PlantSecKB2 将会为研究团队提供一个有效的全基因组比较分析，为进一步探索真菌分泌蛋白在生物燃料、环境修复和植物人类的病原真菌的防治中的功能发挥作用。

作者贡献

JM 完成数据库，DA 收集物种的生活方式数据，KA 和 GZ 参与研究方法的研究，XJM 和 CC 构思研究。设计数据处理程序。XJM、JM、DA 和 CC 负责分析数据以及论文的撰写。全体作者同意该文章的出版。

致谢

我们非常感谢 Gengkon Lum 博士和 Feng Yu 他们在维护服务器方面做出的工作，Jessica Orr 和 Stephanie Frazier 在分泌蛋白数据的人工获取工作的帮助。

参考文献

Alfaro M., Oguiza J.A., Ramírez L. et al. 2014, Comparative analysis of secretomes in basidiomycete fungi, J Proteomics, 102C: 28-43

Bendtsen J.D., Jensen L.J., Blom N. et al. 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. et al. 2004b, Improved prediction of signal peptides: SignalP 3.0, J Mol Biol, 340: 783-795

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

Braaksma M., Martens-Uzunova E.S., et al. 2010, An inventory of the Aspergillus niger secretome by combining in silico predictions with shotgun proteomics data, BMC Genomics, 19: 584

Brown N.A., Antoniw J., and Hammond-Kosack K.E., 2012, The predicted secretome of the plant pathogenic fungus Fusarium graminearum: a refined comparative analysis, PLoS One, 7: e33731

Caccia D., Dugo M., Callari M., et al. (2013) Bioinformatics tools for secretome analysis, Biochim. Biophys. Acta., 1834: 2442-2453

Choi J., Park J., Kim D., et al. 2010, Fungal secretome database: integrated platform for annotation of fungal secretomes, BMC Genomics, 11: 105

Cortázar A.R., Aransay A.M., Alfaro M., et al. 2014, SECRETOOL: integrated secretome analysis tool for fungi, Amino Acids, 46: 471-473

de Castro E., Sigrist C.J., Gattiker A., et al. 2001 ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins, Nucleic Acids Res., 34: W362-365

Do Vale L.H., Gómez-Mendoza D.P., Kim M.S., et al. 2012, Secretome analysis of the fungus Trichoderma harzianum grown on cellulose, Proteomics, 12: 2716-2728

Druzhinina I.S., Shelest E., and Kubicek C.P., 2012, Novel traits of Trichoderma predicted through the analysis of its secretome, FEMS Microbiol Lett., 337: 1-9

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

Ene I.V., Heilmann C.J., Sorgo A.G., et al. (2012) Carbon source-induced reprogramming of the cell wall proteome and secretome modulates the adherence and drug resistance of the fungal pathogen Candida albicans,

Proteomics, 12: 3164-3179

Girard V., Dieryckx C., Job C. et al. 2013, Secretomes: the fungal strike force, Proteomics, 13: 597-608

Horton P., Park K.-J., Obayashi T., et al. 2007, WoLF PSORT: protein localization predictor, Nucleic Acids Res., 35: W585-587

Jung Y.H., Jeong S.H., Kim S.H., et al. 2012, Secretome analysis of Magnaporthe oryzae using in vitro systems, Proteomics, 12: 878-900

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: W429-432

Krogh A., Larsson B., von Heijne G., et al. 2001, Predicting transmembrane protein topology with a hidden Markov

model: Application to complete genomes, J. Mol. Biol., 305: 567-580

Lange L., Bech L., Busk P.K., et al. 2012, The importance of fungi and of mycology for a global development of the bioeconomy, IMA Fungus, 3: 87-92

Lee S.A., Wormsley S., Kamoun S., et al. 2003, An analysis of the Candida albicans genome database for soluble secreted proteins using computer-based prediction algorithms, Yeast, 20: 595-610

Lowe R.G., and Howlett B.J., 2012, Indifferent, affectionate, or deceitful: lifestyles and secretomes of fungi, PLoS pathogens, 8: e1002515

Lum G., and Min X.J., 2011, FunSecKB: the fungal secretome knowledgebase, Database (Oxford), 2011, doi: 10.1093/database/bar001

Lum G., and MinX.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., Meinken J., Orr J., et al. 2014, PlantSecKB: the plant secretome and subcellular proteome knowledgebase. Comput. Mole. Biol., 4: 1-17 Martinez D., Challacombe J., Morgenstern I., et al. 2009, Genome, transcriptome, and secretome analysis of wood decay fungus Postia placenta supports unique mechanisms of lignocellulose conversion, Proc Natl Acad Sci U S A,106: 1954-1959

Martinez D., Larrondo L.F., Putnam N., et al. 2004, Genome sequence of the lignocellulose degrading fungus Phanerochaete chrysosporium strain RP78, Nat Biotechnol. 22: 695-700

McCarthy F.M., Wang N., Magee G.B., et al. 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: 1-7

Melhem H., Min X.J., and Butler G., 2013, The impact of SignalP 4.0 on the prediction of secreted proteins. 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

Morais do Amaral A., Antoniw J., Rudd J.J., et al. 2012, Defining the predicted protein secretome of the fungal wheat leaf pathogen Mycosphaerella graminicola, PLoS One. 7: e49904

Mueller O., Kahmann R., Aguilar G., et al. 2008, The secretome of the maize pathogen Ustilago maydis, Fungal Genet. Biol., 1: S63-S70

Murphy C., Powlowski J., Wu M., et al. 2011, Curation of characterized glycoside hydrolases of fungal origin, Database (Oxford). 2011,

Paper J.M., Scott-Craig J.S., Adhikari N.D., et al. 2007, Comparative proteomics of extracellular proteins in vitro and in planta from the pathogenic fungus Fusarium graminearum, Proteomics, 7: 3171-3183

PetersenT.N., Brunak S., von Heijne G., et al. 2011, SignalP 4.0: discriminating signal peptides from transmembrane regions, Nature Methods, 8: 785-786

Peterson R., Grinyer J., and Nevalainen H., 2011, Secretome of the coprophilous fungus Doratomyces stemonitis C8, isolated from koala feces, Appl. Environ. Microbiol., 77: 3793-3801

Poisson G., Chauve C., Chen X., et al. 2007, FragAnchor a large scale all Eukaryota predictor of Glycosylphosphatidylinositol-anchor in protein sequences by qualitative scoring, Genomics Proteomics Bioinform., 5: 121-130

Powers-Fletcher M.V., Jambunathan K., Brewer J.L., et al. 2011, Impact of the lectin chaperone calnexin on the stress response, virulence and proteolytic secretome of the fungal pathogen Aspergillus fumigatus,

PLoS One, 6: e28865

Ribeiro D.A., Cota J., Alvarez T.M., et al. 2012, The Penicillium echinulatum secretome on sugar cane bagasse, PloS One, 7: e50571

Salvachúa D., Martínez A.T., Tien M, et al. 2013, Differential proteomic analysis of the secretome of Irpex lacteus and other white-rot fungi during wheat straw pretreatment, Biotechnol. Biofuels. 6: 115

Sigrist C.J.A., Cerutti L., de Casro E., et al. 2010, PROSITE, a protein domain database for functional characterization and annotation, Nucleic Acids Res., 38: 161-166

The UniProt Consortium, 2014, Activities at the Universal Protein Resource (UniProt). Nucleic Acids Res., 42:D191-198

Tjalsma H., Bolhuis A., Jongbloed J.D., et al. 2000, Signal peptide-dependent protein transport in Bacillus subtilis: a genome-based survey of the secretome, Microbiol. Mol. Biol. Rev., 64: 515-547

Tsang A., Butler G., Powlowski J., et al. 2009, Analytical and computational approaches to define the Aspergillus niger secretome, Fungal Genetics Biol., 46:S153-160

Weber S.S., Parente A.F.A., Borges C.L., et al. 2012, Analysis of the secretomes of Paracoccidioides mycelia and yeast cells, PloS One, 7: e52470

Wymelenberg A.V., Sabat G., Martinez D., et al. 2005, The Phanerochaete chrysosporium secretome: database predictions and initial mass spectrometry peptide identifications in cellulose-grown medium, J.Biotechnol., 118: 17-34

Yajima W., and Kav N.N., 2006, The proteome of the phytopathogenic fungus Sclerotinia sclerotiorum, Proteomics, 6: 5995-600