10 PubChem

PubChem是NIH设立分子图书馆以及开展研究小分子化学、结构和生物学特性工作的基础。三个Entrez数据库——PCSubstance、PCCompound和PCBioAssay收录了所有的相关信息。这三个数据库共收录有将近4100万条小分子记录和1900万种结构。其中750,000条记录都在PubChem中收录的1200种生物检测方法中的至少一种生物检测试验中具有活性。PubChem不仅与PubMed、PMC等Entrez数据库有链接，还与Entrez Structure和Entrez Protein有链接，这样就将基因组水平的生物大分子与细胞代谢水平的小分子联系起来了。用户可以使用文本在PubChem数据库中进行搜索，也可以使用各种格式的化学分子式或化学结构进行搜索。

更多信息

用户可以在NCBI Bookshelf中找到NCBI手册，该手册详细的介绍了NCBI中的各种资源。在NCBI的主页上还有“教育（Education）”链接，其中有多个教程可供用户学习。用户还可以通过网站地图了解NCBI中的各种资源。在“About NCBI”网页上还有生物信息学入门和其它补充资源。NCBI还设有面向用户的服务人员回答各种问题，用户可以发邮件到info@ncbi.nlm.nih.gov向他们提问。www.ncbi.nlm.nih.gov/About/newsletter.html网页上还有NCBI最新的更新信息。用户还可以到www.ncbi.nlm.nih.gov/Sitemap/Summary/email_lists.html订阅各种更新信息邮件。NCBI现在还设有RSS服务。

原文检索：*Nucleic Acids Research*, 2009, Vol. 37, Database issue D5–D15

YORK /编译

二 蛋白序列数据库

全球蛋白资源数据库UniProt（2009更新）

图片来源：UniProt

UniProt是一个集中收录蛋白质资源并能与其它资源相互联系的数据库，也是目前为止收录蛋白质序列目录最广泛、功能注释最全面的一个数据库。UniProt是由欧洲生物信息学研究所（European Bioinformatics Institute）、美国蛋白质信息资源（Prontein Information Resource）以及瑞士生物信息研究所（Swiss Institute of Bioinformatics）等机构共同组成的UniProt协会（UniProt Consortium）编辑、制作的一个信息资源，旨在为从事现代生物研究的科研人员提供一个有关蛋白质序列及其相关功能方面的广泛的、高质量的并可免费使用的共享数据库。

UniProt是一个向所有使用者免费开放的数据库，全球科研人员都可以登陆网站www.uniprot.org浏览并下载这些资料。借助它，科研人员可以对目的蛋白进行交互式分析或特定的分析。

1 UniProt数据库的构成

UniProt数据库由UniProt知识库（UniProtKB）、UniProt档案（UniParc）、UniProt参考资料库（UniRef）以及UniProt元基因组学与环境微生物序列数据库（UniMES）构成。

1.1 UniProt知识库（UniProtKB）

UniProt知识库是一个专家级的数据库，它可以通过与其它资源进行交互查找的方式为用户提供一个有关目的蛋白质的全面的综合信息。

UniProtKB包括两个组成部分：UniProtKB/Swiss-Prot与UniProtKB/TrEMBL。

1.1.1 UniProtKB/Swiss-Prot

UniProtKB/Swiss-Prot主要收录人工注释的序列及其相关文献信息和经过计算机辅助分析的序列。这些注释都是由专业的生物学家给出的，准确性无需置疑。在UniProtKB中，注释包括对蛋白质功能、酶学特性、具有生物学意义的相关结构域及位点、翻译后修饰情况、亚细胞定位、组织特异性、发育阶段特异性、结构、相互作用、剪接异构体、相关疾病信息的注释等等。注释的另一个重要工作就是对同一蛋白的所有相关报道进行归纳、总结。对蛋白质序列进行仔细检查之后，注释人员还会将相关参考序列、剪接变异体、基因变异体和疾病相关信息全都整合起来，而且不同序列间有任何的差异也会标示出来。注释人员还会将蛋白质数据与其它核酸数据库、物种特异性数据库、结构域数据库、家族遗传史或疾病资料数据库进行交叉参考。

1.1.2 UniProtKB/TrEMBL

UniProtKB/TrEMBL收录的则是高质量的经计算机分析后进行自动注释和分类的序列。计算机辅助注释使用的是Spearmint规则，而人工注释依据的则是蛋白质家族规则，包括HAMAP家族规则（HAMAP family rules）、RuleBase规则、PIRSF分类命名规则以及位点规则。UniProtKB/TrEMBL还收录了所有EMBL-Bank/ GenBank/DDBJ核酸序列数据库中的编码序列的翻译后蛋白质序列和来自拟南芥信息资源库（TAIR）、SGD和人类Ensembl数据库中序列的翻译后蛋白质序列。

其中，研究人员排除了诸如EMBL-Bank/ GenBank/DDBJ数据库中编码小片段的序列、人工合成的序列、大部分非胚系免疫球蛋白序列、大部分T细胞受体序列、大部分专利序列和一些高度过表达的序列。这些选择的记录都是经过大量人工注释的，然后根据注释的情况收入UniProtKB/Swiss-Prot数据库。

1.2 UniProt档案（UniParc）

UniProt档案则是关于蛋白质序列的全面数据库，它储存了大量的蛋白质序列资源，反映了所有蛋白质序列的历史。

UniParc是储存序列的数据库，同时也是最全面的能反映所有蛋白质序列历史的数据库。UniParc收录了不同数据库来源的所有的最新蛋白质序列和修订过的蛋白质序列，因此可以保证数据收录的全面性。

UniParc数据库收录的资源

UniProtKB数库据	NCBI的RefSeq数据库
EMBL-Bank/DDBJ/GenBank这些核酸数据库中核酸序列的翻译后序列	模式生物数据库FlyBase
Ensembl数据库中真核生物基因组数据	SGD
H-邀请数据库（H-Inv）	TAIR
脊椎动物基因组注释数据库（VEGA）	WormBase
IPI数据库	TROME
蛋白质研究基金会数据库（PRF）	美国、欧洲、韩国、日本专利局中的数据
蛋白质数据库（PDB）	

为了避免出现冗余数据，UniParc将所有完全一样的序列都合并成了一条记录，而不论这些数据是否来自同一物种。UniParc还会收录每天最新的数据和修改过的数据，并交叉参考这些数据，及时对UniParc中的数据做出修订。UniParc中每一条记录包含的基本信息包括标识符、序列、循环冗余校验码、来源数据库中的检索号、版本号、时间印记。如果UniParc中的记录没有收录在UniProtKB中，那么这个基因可能是假基因。此外，除了给出每一条记录在来源数据库中的检索号之外还会给出这条记录在来源数据库中的状态，例如是仍然存在或者是已经被删除，也会给出NCBI GI号和TaxId号。UniParc中的记录都是没有注释的，因为蛋白质只有在指定的条件下才能够进行注释。例如，序列完全相同的蛋白质如果属于不同的物种、组织或不同的发育阶段，其功能都有可能完全不同。

1.3 UniProt参考资料库（UniRef）

UniProt参考资料库可以通过序列同一性对最相近的序列进行归并，加快搜索速度。

UniRef对来自UniProtKB的各种数据包括各种剪接变异体进行了分类汇总，还从UniParc中选取了一些数据以求能完整的、没有遗漏的收录所有数据，同时也保证没有冗余数据，该数据库的同一性（identity）分为三个级别：100%、90%和50%。

UniRef里的数据是按照级别来分类的，在UniRef数据库的每一个同一性级别中，每一条序列只会属于其中的一个聚类，这条序列在其它的同一性级别中也只会有一条父集（parent cluster）序列和子集（child cluster）序列。UniRef100数据库将相同的序列数据和亚片段数据整合在一起，使用一个检索入口进行检索。

UniRef90数据库建立在UniRef100数据库的基础之上，而UniRef50数据库又是以UniRef90为基础。UniRef100、UniRef90和UniRef50这三个数据库的数据量分别减少10%、40%和70%。每一个聚类记录都包含下列信息：数据来源、蛋白质名称、分类学信息（但只会举一个蛋白质为代表）、聚类下条目数等。UniRef100是目前最全面的非冗余蛋白质序列数据库。UniRef90和UniRef50数据量有所减少是为了能更快地进行序列相似性搜索以减少结果的误差。UniRef现在已广泛用于自动基因组注释、蛋白质家族分类、系统生物学、结构基因组学、系统发生分析、质谱分析等各个研究领域。UniRef中的聚类信息是会随着UniProtKB的更新而同步更新的。

1.4 UniProt元基因组学与环境微生物序列数据库（UniMES）

UniProt元基因组学与环境微生物序列数据库是为不断发展壮大的元基因组学研究领域服务的。

UniProt Knowledgebase中的Swiss-Prot和TrEMBL两个数据库包含了分类学信息明确的序列数据。不过，不断增多的元基因组数据迫使人们需要另外再建一个数据库，即UniMES。目前，UniMES收录了来自全球海洋取样考察计划（GOS）得来的数据，而GOS以前则将数据上传至国际核酸序列数据库协作体（INSDC）。GOS的数据包含有大约2500万条DNA序列，估计可以编码大约600万种蛋白质，这些序列都是来自于海洋微生物。UniMES将这些可能的蛋白质序列和InterPro数据库自动分类、整理后的序列资源结

合起来，成为了目前唯一能提供全球海洋取样考察计划获得的基因组信息数据库，同时它还是免费使用的。UniMES中的数据没有收录在UniProtKB和UniRef中，但UniParc中有收录。UniMES中的数据以FASTA形式储存，可以从FTP服务器上免费下载。

2 新进展

新的UniProt网站

UniProt协会发布了他们最新的官方网站，该网站有新的界面，新的搜索引擎，有更多的新选项方便大家使用。之前的镜像网站（www.ebi.uniprot.org、www.expasy.uniprot.org、www.pir.uniprot.org、www.expasy.org）则都被取消了。UniProt还提供使用本体术语扩展查询结果的功能。

同时，UniProt对序列相似性搜索、多序列比对、分批处理和数据库标识符作图工具这些最常用的生物信息学工具也都进行了简化。

用户还可以在www.uniprot.org/help/technical网站上通过简单的HTTP（REST）操作进行编程查询操作。

网站上除了现有的文件格式（例如纯文本格式、FASTA格式、UniProtKB 中的XML格式等）之外还提供了可配置的空位分隔格式（tab-delimited）、RSS格式和GFF格式供用户下载资料，同时，所有的文件都有RDF格式（www.w3.org/RDF/）和在语义网（Semantic Web）上使用的W3C格式下载。更多信息请浏览www.uniprot.org/help/。

3 UniProtKB附加的蛋白质文献信息

UniProt一直致力于将UniProtKB注释蛋白质时引用的文献等信息整合到UniProt中以供用户参考。目前，有将近218,000条PubMed的文献被引用来注释UniProtKB中将近410万条序列，而这些文献中有66%都被收录到UniProtKB/Swiss-Prot中。其它诸如Entrez Gene数据库、模式生物数据库（MOD）、SGD、MGI等公共数据库也都为每条基因或蛋白记录提供引用文献信息。对于那些在不同数据库中都被注释过的基因来说，每一个数据库都会根据自己的特点来有选择的引用相关文献进行注释。因此，将各种不同的数据库文献资源都整合到UniProtKB非常有必要。UniProt现在已经将收录人类、小鼠、酵母和其它物种基因或蛋白质信息的5个外部数据库的引用文献信息整合进来了，这些外部数据库包括：Entrez Gene里的GeneRIF数据库（www.ncbi.nlm.nih.gov/projects/GeneRif）、SGD（www.yeastgenome.org）、MGI（www.informatics.jax.org）、GAD（geneticassociationdn.nih.gov）以及PDB（www.rcsb.org/pdb/）。

上述5个外部数据库中共整合了约244,000条来自PubMed同时UniProtKB中还不曾收录的引用文献，这些文献涵盖了UniProtKB中约110,000条记录。其它额外的文献记录都直接链接到UniProt蛋白质查询网页上了。UniProt还将继续从其它MOD数据库和蛋白质功能数据库中发掘更多的文献资料补充到UniProtKB中。这些补充的文献资料不仅有利于对UniProtKB中的记录进行注释，同时也有利于帮助用户发掘出更多他们感兴趣的蛋白的资料。

4 格式改变

UniProt格式的改变是为了改善资料的一致性（consistency）和可用性（usability）。UniProt建议用户密切关注它的newsfeeds，以充分利用这些改变带来的便利。最近几个月来发生的以及在未来几个月里将要发生的格式变化，请浏览www. uniprot.org/。

5 UniProtKB注释

UniProtKB包括两个部分：Swiss-Prot和TrEMBL。

UniProtKB/Swiss-Prot包含人工注释的记录，及其相关参考文献和计算机辅助分析信息。人工注释信息包括蛋白质序列和试验证据或计算机预测信息。还有许多生物学专家不断的对这些数据进行完善和补充。对UniProtKB/Swiss-Prot中的记录进行注释的工作可以分为两个部分：模式生物的注释以及横向注释。

5.1 模式生物的注释

UniProtKB/Swiss-Prot对许多物种的蛋白质进行了注释，但主要还是集中在对分类清楚的模式生物蛋白的注释上，因为只有这样才能保证对每一个蛋白质家族的“代表”做出高质量的注释。

被注释的模式生物包括

人类及其它哺乳动物、相关数据库HPI	病毒
细菌和古细菌、相关数据库HAMAP	毒素、相关数据库Tox-Prot
植物、相关数据库PPAP	果蝇、非洲蟾蜍、斑马鱼和秀丽隐杆线虫
真菌、相关数据库FPAP	

5.2 横向注释（Transversal annotation）

横向注释主要关注存在于所有物种中的普遍现象，例如翻译后修饰（PTM）、蛋白质结构信息、蛋白间相互作用等。了解更多内容请浏览www.uniprot.org/help/projects

6 第一张完整的人类基因组草图

最近收录在UniProtKB/Swiss-Prot中的注释成果就是完成了对第一张完整的人类基因组草图的注释工作。这意味着在UniProt 14.1版本中可以查询到所有已知人类蛋白编码基因的人工注释结果。在该版本发布时共收录了20,325条记录，其中超过1/3的记录还都收录有因可变剪接、可变启动子或可变翻译起始位点造成的异构体序列。因此该版本收录的蛋白质序列将近有34,000条，还收录有大约46,000个单氨基酸多态性（SAP）和60,000个PTM，这些大部分都与人类疾病有关。

这已经不是UniProtKB/Swiss-Prot第一次对一个模式生物全蛋白质组进行注释了，还曾经对大肠杆菌和酿酒酵母的全蛋白质组进行过注释，将来还将对更多的如拟南芥（*Arabidopsis thaliana*）、枯草芽孢杆菌（*Bacillus subtilis*）、盘基网柄菌（*Dictyostelium discoideum*）、小鼠、水稻、金黄色葡萄球菌（*Staphylococcus aureus*）、裂殖酵母（*Schizosaccharomyces pombe*）等模式生物的全蛋白质组进行注释。不过可能这些工作不如对人类蛋白质组进行注释那么重要。

UniProt第一次能够向全世界清晰的展示一幅完整的（虽然不尽完美，但至少UniProt认为是完整的）人类蛋白质图画。能在分子水平对人类蛋白质有一个总体的了解是全世界生物学家共同的最终目标，UniProt希望这项工作能帮助科学家早日实现这个愿望。不过还有许多难题等待着人们去解决。人们还将继续收录新发现的人类蛋白质，不断整理、修改已经收录的记录，收录更多的剪接变异体，发现更多的PTM。总之，就是不断完善人类蛋白质组数据库。分子水平的研究成果也应该放到生理水平，例如亚细胞定位、组织表达和蛋白相互作用等水平去进行验证。

原文检索：*Nucleic Acids Research*, 2009, Vol. 37, Database issue D169–D174

YORK /编译

1 Metagenomic

Metagenomic是元基因组学也有译作宏基因组学，是对环境样品中微生物群体基因组进行分析的一种方法。

2 HAMAP：High-quality Automated and Manual Annotation of microbial Proteome

HAMAP是高品质微生物蛋白质组的自动化和手动的注释。该计划旨在将人工注释方法与计算机注释方法结合起来，力求在保证注释质量的前提下，提高注释速度。

3 Splice variant

Splice variant是指剪接变异体。真核生物的基因有外显子与内含子两部分。真核生物基因有外显子、内含子的一个结果就是其基因产物可能有不同的长度，因为并非所有的外显子都包含在最终的mRNA中（包含在mRNA内的外显子的排列顺序没有改变）。由于mRNA的编辑产生了不同的多肽，进而形成不同蛋白质，这些蛋白质就互称为剪接变异体或者可变剪切形式。

4 H-Invitational Database

H-Invitational Database即H-邀请数据库，是一个有关人类基因及转录体的全面数据库。它可以提供基因注释、基因结构、可变剪接异构体、非编码功能RNA、蛋白质功能、蛋白质功能结构域、蛋白质亚细胞定位、蛋白质三维结构、代谢通路、遗传多态性、基因表达谱、分子进化、蛋白间相互作用、基因家族、相关疾病等多种信息。

5. 语义网

语义网是Semantic Web的中文名称。语义网就是能够根据语义进行判断的网络。简单地说，语义网是一种能理解人类语言的智能网络，它不但能够理解人类的语言，而且还可以使人与电脑之间的交流变得像人与人之间交流一样轻松。语义网是对未来网络的一个设想，在这样的网络中，信息都被赋予了明确的含义，机器能够自动地处理和集成网上可用的信息。语义网使用XML来定义定制的标签格式以及用RDF的灵活性来表达数据，下一步需要的就是一种Ontology的网络语言，比如OWL来描述网络文档中的术语的明确含义和它们之间的关系。

语义网与万维网的区别：

目前人们所使用的万维网，实际上是一个存储和共享图像、文本的媒介，电脑所能看到的只是一堆文字或图像，对其内容无法进行识别。万维网中的信息，如果要让电脑进行处理的话，就必须首先将这些信息加工成计算机可以理解的原始信息后才能进行处理，这是相当麻烦的事情。而语义网的建立则将事情变得简单得多。

语义网是对万维网本质的变革，它的主要开发任务是使数据更加便于电脑进行处理和查找。其最终目标是让用户变成全能的上帝，对因特网上的海量资源达到几乎无所不知的程度，计算机可以在这些资源中找到用户所需要的信息，从而将万维网中一个个现存的信息孤岛，发展成一个巨大的数据库。

语义网将使人类从搜索相关网页的繁重劳动中解放出来。因为网中的计算机能利用自己的智能软件，在搜索数以万计的网页时，通过“智能代理”从中筛选出相关的有用信息。而不像现在的万维网，只给你罗列出数以万计的无用搜索结果。例如，在进行在线登记参加会议时，会议主办方在网站上列出了时间、地点，以及附近宾馆的打折信息。如果使用万维网的话，此时你必须上网查看时间表，并进行拷贝和粘贴，然后打电话或在线预订机票和宾馆等。但假如使用的是语义网，那么一切都变得很简单了，此时安装在你计算机上的软件会自动替你完成上述步骤，你所做的仅仅是用鼠标按几个按钮而已。在浏览新闻时，语义网将给每一篇新闻报道贴上标签，分门别类的详细描述哪句是作者、哪句是导语、哪句是标题。这样，如果你在搜索引擎里输入“老舍的作品”，你就可以轻松找到老舍的作品，而不是关于他的文章。

总之，语义网是一种更丰富多彩、更个性化的网络，你可以给予其高度信任，让它帮助你滤掉你所不喜欢的内容，使得网络更像是你自己的网络。