中国科学: 生命科学 2024 年 第 54 卷 第 6 期: 1101 ~ 1108
SCIENTIA SINICA Vitae lifecn.scichina.com
《中国科学》杂志社 SCIENCE CHINA PRESS
评 述 生物样本资源专辑

蛋白质组科学数据库建设及应用

刘祎, 徐小放, 马洁, 陈涛*, 朱云平*

军事科学院军事医学研究院生命组学研究所, 国家蛋白质科学中心(北京), 北京蛋白质组研究中心, 蛋白质组学国家重点实验室, 北京 102206
* 联系人, E-mail: taochen1019@163.com; zhuyunping@gmail.com

收稿日期: 2023-09-15; 接受日期: 2023-11-22; 网络版发表日期: 2024-04-30

摘要 蛋白质组技术的发展及推广应用使蛋白质组数据共享的需要日益迫切, 在数据共享理念的支撑下, 蛋白质组数据库不断涌现. 本文首先介绍了蛋白质组科学数据库的建设情况, 重点阐述了国际蛋白质组数据联盟以及成员的现状, 详细介绍了中国的蛋白质组数据库iProX, 然后简要介绍了当前蛋白质组学数据库应用的四个方向, 最后对该领域存在的挑战和机遇进行了展望.

关键词 蛋白质组学, 数据库, 数据共享

随着基因组研究的不断发展和人类基因组测序的完成, 蛋白质组学蓬勃兴起, 成为后基因组时代最重要的研究领域之一. 蛋白质组学从整体的角度研究细胞、组织、器官或生物体内蛋白质的组成成分、表达水平、修饰状态、相互作用及动态变化, 系统地揭示生命活动规律, 探索疾病的发生发展机制, 为疾病的预防、诊断和治疗提供理论依据和解决途径. 2001年, 22位国际知名科学家发起成立国际人类蛋白质组组织(Human Proteome Organization, HUPO). 2002年, 在HUPO组织召开的第一次研讨会上, 与会科学家提出了“人类蛋白质组计划”(Human Proteome Project, HPP), 这是21世纪第一个重大国际合作大科学工程. 最早启动的是美国的人类血浆蛋白质组学计划和中国的人类肝脏蛋白质组学计划. 2010年底, HPP共启动了11项分项计划. 同年, HUPO再次正式启动了HPP计划, 2020年完成涵盖超过90%的人类蛋白质. 2022年, 中国科学家领衔发起并主导“人体蛋白质组导航计划”(Proteomic Navigator of the Human Body, π-HuB).

随着人类蛋白质组计划的全面发展, 蛋白质组实验数据源源不断地产出[1], 蛋白质组数据共享的需求日益迫切. 基因组研究的快速发展极大地受益于人类基因组计划的开放数据共享[2], 在蛋白质组研究不断深入的20年多年里, 数据的开放共享也逐渐成为蛋白质组学领域的共识. 在阿姆斯特丹原则[3]及其后续蛋白质组数据质量控制标准[4,5]的推动下, 支持快速共享和开放共享蛋白质组数据的政策和指南在蛋白质组学界得到推广和实施. 与此同时, 以蛋白质组数据标准组织(Human Proteome Organization Proteomics Standards Initiative, HUPO-PSI)为代表的团体不断推进蛋白质组数据标准和信息指南的制定和发布, 如数据交换格式、受控词汇表[6,7]和报告指南(即The Minimum Information about a Proteomics Experiment, MIAPE[8])等都取得了重大进展. 近10年内, HUPO-PSI[9]制定并发布的蛋白质组学数据标准文件格式包括表示质谱原始数

引用格式: 刘祎, 徐小放, 马洁, 等. 蛋白质组科学数据库建设及应用. 中国科学: 生命科学, 2024, 54: 1101–1108
Liu Y, Xu X F, Ma J, et al. Current status and applications of proteomic databases (in Chinese). Sci Sin Vitae, 2024, 54: 1101–1108, doi: 10.1360/SSV-2023-0141

据的mzML[10]标准、设计用于报告肽段鉴定结果的mzIdentML[11]标准、用于报告定量结果的mzQuantML[12]标准和基于文本的格式mzTab[13]. 所有这些标准格式及其转换工具[14~17]都有助于促进蛋白质组学数据的共享.

蛋白质组科学数据库是实现蛋白质组数据共享必不可少的条件. 蛋白质组科学数据库对来自不同国家地区不同实验室产出的实验数据进行长期而稳定的存储, 并对蛋白质组研究人员提供实验数据的评估、重分析、比较和利用等方面的支持, 进一步推动蛋白质组研究的蓬勃发展.

1 蛋白质组科学数据库的建设

当前, 蛋白质组学领域主要采用基于质谱的检测方法. 串联质谱的进步使得在基于质谱的检测中鉴定数十万种蛋白质成为可能[18]. 可以说, 质谱数据是当前蛋白质组学的基础. 由于质谱数据与其他类型的数据(如表达量数据、序列数据)有较大区别, 根据数据库是否存储质谱数据, 本文将涉及的数据库分为蛋白质组质谱实验数据库与其他数据库, 如表1所示. 需要注意的是, 目前存储质谱原始数据的数据库不仅存储质谱实验的原始数据, 而且也会涉及后续的鉴定结果等其他信息. 其他数据库指根据质谱分析结果整合构建的二级数据库. 蛋白质组学分析流程中还会涉及一些与质谱数据并不直接相关的数据库(如蛋白质序列数据库), 本文不涉及该类型的数据库, 同时也不包括已无法访问或长久未更新的数据库.

1.1 蛋白质组质谱实验数据库

2012年以来, 国际蛋白质组数据联盟(ProteomeXchange, PX, http://www.proteomexchange.org)[1,19,20]已经在全球范围内对蛋白质组质谱实验数据的提交和发布制定了国际标准. PX联盟致力于协调和稳定蛋白质组数据的共享, 支持并实施HUPO-PSI的数据格式. 截至2019年6月底, PX包含14100个数据集, 其中8638个数据集(61%)已经公开. 数据集每年都在大幅增加, 来自于50多个不同的国家, 具有强大的国际影响力. PX目前包括PRIDE, PeptideAtlas, MassIVE, jPOST, iProX和Panorama共6个正式成员, 其中PRIDE和PeptideAtlas

表 1 基于质谱的蛋白质组学数据库列表

Table 1 List of mass spectrometry-based proteomics databases

类别	数据库名称	数据库链接	描述
蛋白质组质谱实验数据库	PRIDE	http://www.ebi.ac.uk/pride/archive	PX联盟成员, 支持质谱原始文件的提交
	PeptideAtlas	http://www.peptideatlas.org	PX联盟成员, 支持质谱原始文件的提交和数据的统一分析
	MassIVE	https://massive.ucsd.edu/ProteoSAFe/static/massive.jsp	PX联盟成员, 支持质谱原始文件的提交
	jPOST	https://repository.jpostdb.org/	PX联盟成员, 支持质谱原始文件的提交
	iProX	https://www.iProX.cn/	PX联盟成员, 支持质谱原始文件的提交
	Panorama	https://panoramaweb.org/project/home/begin.view?	PX联盟成员, 支持质谱原始文件的提交, 主要面向靶向质谱数据
	Tranche	https://proteomecommons.org/tranche/	支持质谱原始文件的提交, 需要使用客户端
其他数据库	GPMDB	http://www.thegpm.org/	支持质谱数据的分析
	PepQuery	http://www.pepquery.org/index.html	支持肽段的在线搜索
	PaxDb	https://pax-db.org/	收集蛋白的表达量数据
	PPD	http://www.plasmaproteomedatabase.org/	收集血液蛋白质组数据
	PPDB	http://ppdb.tc.cornell.edu/	收集植物蛋白质组数据
	RPD	http://www.info.chi-biotech.cc/	收集水稻蛋白质组数据
	MMPdb	https://mmpdb.eeob.iastate.edu	收集线粒体蛋白质组数据
	Phospho.ELM	http://phospho.elm.eu.org/	收集蛋白质组磷酸化数据
	PhosphoSitePlus	https://www.phosphosite.org/homeAction.action	收集蛋白质组翻译后修饰数据
	PHOSIDA	http://141.61.102.18/phosida/index.aspx	收集蛋白质组翻译后修饰数据

为该联盟的发起者, iProX由本团队开发并负责运行维护. 所有PX成员共享数据集标识符空间(PXD或者RPXD标识符), 通过标识符进行发布和追踪.

PRIDE[17,21]由欧洲生物信息学研究所(European Molecular Biology Laboratory European Bioinformatics Institute, EMBL-EBI)创建, 是世界上最大的基于质谱的蛋白质组学数据库. 截至2023年2月, PRIDE已经储存了约2万个公开数据集, 累积了约1.5亿蛋白质证据(protein evidence)、4.2亿肽段证据(peptide evidence)和5.8亿谱图证据(PSM evidence). PRIDE在HPP中发挥了重要作用.

PeptideAtlas[18,22]是一个存储质谱原始数据和基于质谱数据各种分析结果的数据库. 其中的原始数据会基于标准的串联质谱蛋白质组分析流程(trans-proteomic pipeline, TPP)定期进行分析, 分析结果会通过网络共享. PeptideAtlas可以帮助研究者规划有针对性的蛋白质组学实验、改进基因组注释并支持数据挖掘项目.

MassIVE于2014年加入PX联盟, 同样是一个用于存储质谱原始数据的数据库. 与此同时, MassIVE还支持数据的在线重分析(如基于MS-GF+的蛋白鉴定). 截至2023年2月, MassIVE已经累积了489.83 TB的数据, 包括13463个公开数据集.

jPOST[23~25]于2014年加入PX联盟, 由日本创建的一个共享蛋白质组原始/分析数据的数据存储库. 用户可以通过JPOSTTrepo使用唯一标识号共享和发布数据. 目前已经累积了1060个公开数据集, 数据量达到64.3 TB.

iProX[26,27]是位于中国的综合性蛋白质组数据库, 于2017年加入PX联盟, 旨在支持以中国研究团队为主的蛋白质组学研究数据的全球共享. iProX实现了蛋白质组学实验相关原始数据、分析结果和标准化元数据的高效汇集和共享, 为学术界提供了一个蛋白质组数据的发布及共享平台, iProX系统在全面兼容PX联盟规则的同时也研发了一系列创新性功能模块, 包括特有的结构化数据管理模式、用户自定义关联数据功能, 以及标准化-自动化的数据审核和发布功能等. 同时, iProX系统依托国家蛋白质科学中心(北京)的设施, 建立了组学大数据云平台的基础环境, 并解决了网络数据传输瓶颈的技术问题, 形成了多地(北京、广州)协同工作的模式, 在有效地进行蛋白质组大数据的汇集、共享和发布的同时保证数据的安全, 并已在国内多个生命科学研究项目中得到了实际应用. 截至2023年6月, 总数据量344.38 TB, 总项目数3883个.

Panorama主要面向靶向质谱数据[28], 于2018年加入PX联盟. 截至目前, 已积累了432个公开项目. Panorama与靶向质谱流程中的Skyline[29]工具有较好的整合, 用户可以直接从Skyline将数据发布到Panorama服务器.

Tranche是一个质谱数据存储库[30], 旨在支持数据和软件的重用和传播. 为了减少数据冗余和实现负载均衡, 它采用对等网络. 它还使用客户端-服务器模型来确保身份验证和可靠性. 上传和下载数据集需要客户端工具. 它有几个重要的特性, 包括发布前加密、数据谱系、数据完整性、不变性和版本控制. 同时, Tranche为PRIDE和PeptideAtlas提供了接口, 以存储和传播大型质谱数据文件.

蛋白质组质谱实验数据库的基本功能是类似的, 都是存储和共享质谱数据. 由于质谱文件往往较大, PX在建立时考虑到人力、资源等因素, 采用了元数据统一管理, 实际数据分散管理的形式. 用户可以根据自身网络条件和习惯选择PX成员之一进行提交. 研究者需要下载公开数据时, 可以统一搜索元数据, 然后进入具体的数据库搜索和下载数据. 与其他数据库相比, iProX支持基于Aspera的高速数据传输, 可以大大节省用户数据上传和下载的时间. 尤其是对国内用户来说, iProX在数据传输速度上有明显的优势. 另一方面, iProX有专人通过邮件、微信和电话等方式提供全流程的支持, 能及时解决用户遇到的问题. 此外, iProX还提供了中英文双语界面.

1.2　其他数据库

其他数据库按存储数据的类型可以分为分析流程相关的数据库、表达量相关的数据库和翻译后修饰相关的数据库.

分析流程相关的数据库包括GPMDB和PepQuery等与质谱数据分析流程密切关联的数据库. GPMDB数据库[31]的构建是为了利用广义蛋白质组学数据元分析(generalized proteomics data meta-analysis, GPM)服务器获得的信息来帮助验证肽段与谱图匹配的准确性, 允许用户快速将实验结果与GPM之前观察到的最佳结果进行比较. 到2023年2月, GPMDB数据涵盖

3695369种蛋白质、22479471种肽段. GPMDB在以染色体为中心的人类蛋白质组(C-HPP)项目中发挥了重要作用. PepQuery数据库提供了对已知或未知肽段的在线搜索, 搜索范围除了该数据库本身包含的48个数据集(超过十亿张二级谱图), 用户也可以指定本地的数据或iProX等数据库的公共数据集[32].

表达量相关的数据库大多以物种或组织类型来划分. PaxDb是[33]一个元数据库, 集成了各种模式生物的绝对蛋白质丰度水平的全生物体数据和组织分辨数据. 它仅从已发表的实验和主要蛋白质组学数据资源(如PRIDE和PeptideAtlas)中导入定量蛋白质组学数据集, 然后分析定量. 到2014年底, 它包括了10482种蛋白质, 143456个肽段和大约2400万张谱图[34]. PaxDb支持跨不同生物群的比较分析, 截至目前, 该数据库已经包含30种真核生物、29种细菌以及2种古菌. 血浆蛋白质组数据库PPD最初于2005年发布, 是人类血浆蛋白质组项目试点计划的一部分[35]. 该数据库目前包含在血清或血浆中检测到的10546种蛋白质的信息, 其中3784种已在两项或多项研究中报告. 该数据库的最新版本纳入了质谱相关的数据, 并启用了名为Plasma Proteome Explorer的批量查询工具, 允许用户针对已知血浆蛋白筛选蛋白质或肽段列表. PPDB是主要面向拟南芥(*Arabidopsis thaliana*)和玉米(*Zea mays*)的植物蛋白质组数据库[36]. 最初PPDB专注于植物质体, 但现在已扩展到整个植物蛋白质组. PPDB中存储了来自质谱分析的实验数据以及有关蛋白质功能、蛋白质特性和亚细胞定位等精选信息. RPMD是第一个针对水稻(*Oryza sativa*)主要器官和组织(包括叶、根、茎、子房、花药和种子)的大规模蛋白质组数据库[37]. 到2019年为止, RPMD包含通过质谱方法表征的17401个蛋白. RPMD还提供转录组水平、翻译组水平和蛋白质组水平的基因定量信息. MMPdb是一个用于促进线粒体蛋白质组比较分析的数据库[38], 包括智人(*Homo sapiens*)、小鼠(*Mus musculus*)、秀丽隐杆线虫(*Caenorhabditis elegans*)、黑腹果蝇(*Drosophila melanogaster*)、卡氏棘阿米巴(*Acanthamoeba castellanii*)和酿酒酵母(*Saccharomyces cerevisiae*)的线粒体蛋白质组数据库.

质谱技术的最新进展使磷酸化蛋白的检测更加有效[39]. Phospho.ELM旨在存储从研究论文和蛋白质组质谱分析中导入的磷酸化数据的数据库[40~42]. 迄今为止, 该网络资源涵盖42914个实例、299个激酶、3657篇参考文献、11224条序列和8698个底物. PhosphoSitePlus旨在收集主要来自人类和小鼠的翻译后修饰数据[43]. PhosphoSitePlus支持两种数据, 包括修饰的氨基酸和周围序列以及与蛋白质功能区域相关的上下游相互作用. PhosphoSitePlus中的大多数位点是使用质谱检测到的, 其余的是根据低通量鉴定方法获得. 目前, PhosphoSitePlus涵盖58127个蛋白质、600115种不同的修饰、588089个高通量质谱位点、30223个低通量位点和28454篇精选论文. PHOSIDA是一个集合了大量高置信度翻译后修饰位点的数据库. 基于质谱的蛋白质组学被用于识别各种物种中的位点[44]. 迄今为止, 该数据库涵盖了80062个N-糖基化、磷酸化或乙酰化位点. 基于极低假阳性率的严格质量标准用于从高分辨率质谱数据中获取这些位点. PHOSIDA包含来自智人(*Homo sapiens*)以及其他物种(包括细菌)的翻译后修饰位点.

一般来说, 对于没有自产数据的研究人员, 可以在章节1.1介绍的这些数据库中搜索与研究问题相关的数据进行分析. 有自产数据的研究人员也可以用公共数据作为参考. 章节1.2介绍的分析流程相关的数据库供研究人员对自己的数据进行进一步挖掘. 表达量相关的数据库和翻译后修饰相关的数据库则供研究人员搜索特定蛋白或肽段的关联信息, 研究人员需要根据物种和数据类型进行选择.

2 蛋白质组科学数据库的应用

蛋白质组学数据类型及其相应数据格式较多. 蛋白质组学数据库(主要是上文中涉及的PX联盟成员)需要存储的主要数据类型是质谱原始数据和对应的鉴定结果. 原始数据的可用性使得对数据集进行全面的重新分析成为可能, 而鉴定结果可以用于可视化和评估给定研究中报告的结果等. 过去数十年间, 数据标准的发展有助于简化数据科学家使用公共蛋白质组学数据的过程.

在一些评论中[45,46], 蛋白质组科学数据库的应用被分为四类: (ⅰ)直接使用(use); (ⅱ) 再使用(reuse); (ⅲ) 再处理(reprocess); (ⅳ) 再利用(repurpose).

直接使用: 蛋白质知识库直接使用本文所述的蛋白质组学数据库的数据进行整合, 如neXtProt[47]知识

库. 这些知识库的使用非常广泛, 很多生命科学领域的研究人员都从这些知识库的数据中受益. 直接使用蛋白质组学数据库中的数据是一种非常有效的应用方法.

再使用: 通过对蛋白质组学数据库中的数据进行再次使用, 从而产生新的知识, 例如, 谱图库是基于蛋白质组学数据库而创建, 谱图库可以在当前流行的质谱数据非依赖(data independent acquisition, DIA)采集模式中发挥重要作用[48]. 通用质谱标识符(Universal Spectrum Identifier, USI)也是一个很好的例子, 对蛋白质组学数据库中的谱图进行重新编码. USI提高了谱图证据的透明度, PX联盟目前已经提供了来自超过30亿张谱图的10亿个USI标识[49]. 此外, 一种通用的数据再使用方法是对来自大量独立数据集的数据进行组合分析, 即所谓的元分析研究(meta analysis), 从而获得新知识, 这些新知识是无法从单个数据集得到的. 但目前在蛋白质组学领域, 由于不同实验室所产生的数据往往难以整合, 这类研究相当稀少[50].

再处理: 因为蛋白质组数据库搜索引擎使用的蛋白质序列数据库越来越精确, 因此对蛋白质组学数据库中的公开数据进行数据库搜索这类的重分析可以提供新的分析结果, 这类重分析与原始实验类似. PeptideAtlas和GPMDB等数据库使用其专用的生物信息学工具和流程定期对质谱原始数据进行重新处理. 在处理过程中, 应用统一的质量控制标准对不同来源的数据进行相互比较. PeptideAtlas的结果被重新组织成数据集, 每个数据集包括来自单一物种蛋白质组(例如, 人类(*Homo sapiens*)、猪(*Sus scrofa domesticus*)、白色念珠菌(*Candida albicans*))或亚蛋白质组(例如, 人类血浆). 类似地, GPMDB也会对用户提供的谱图数据或存储在其他数据库中的原始数据进行再次处理. neXtProt与PeptideAtlas和GPMDB密切合作, 整合了数据再次处理的结果.

再利用则是指基于与当初实验设计有所区别的目的去使用原始数据的情况. 开发新的生信方法时往往会对公共数据库中的数据进行再利用, 例如, 本团队在开发新的基于深度学习的生物标志物算法时[51], 就使用了已公开的肝癌数据[52]. 另外, 蛋白质基因组学方法研究与新的翻译后修饰发现也是蛋白质组科学数据库再次利用的两个应用方向. 典型的应用包括Matic等人[53]的研究和Hahne等人[54]的研究. Matic等人的研究将新修饰添加到标准的Andromeda搜索引擎中, 并鉴定了一批公开的大规模小鼠(*Mus musculus*)组织磷酸化蛋白质组数据. 最终, 他们从79种蛋白质中鉴定出总共88个新的修饰位点. Hahne等人的研究利用一个公开的小鼠(*Mus musculus*)大脑磷酸化蛋白质组数据集, 使用最新开发的Oscore软件重新评估高分辨率串联质谱, 发现了与11种小鼠(*Mus musculus*)蛋白质相对应的23种新的肽的修饰.

目前来看, 第1种(直接使用)和第4种(再利用)情况相对较多, 也较为成熟. 对第1种应用来说, 如何及时整合最新的数据是最大的挑战, 有不少此类数据库在建好一段时间后就放弃更新. 第4种应用需要用户去搜索符合自己实验设计的数据, 但目前各个数据库提供的信息检索较为费力, 用户往往需要进入单个项目仔细阅读项目的信息, 在一定程度上限制了大规模蛋白质组数据的可用性. 第2种(再使用)和第3种(再处理)应用相对稀少, 说明目前蛋白质组学流程的标准化程度还不足, 但这种情况也表明不同来源数据集的整合是一个充满机会和挑战的领域.

3　结论与展望

本文介绍了蛋白质组科学数据库的建设情况. 这些资源不仅提供原始数据和鉴定结果, 还支持前瞻性、高通量的蛋白质组学研究. 此外, 这些蛋白质组数据库还可以为大规模基因组提供注释. 数据库之间共享数据和元数据会变得更加重要. 因此, 蛋白质组学数据库需要一种集成的方法来实现不同数据库之间的数据可访问性. PX联盟提供了一个逻辑集中、物理分布的基础设施, 六个成员PRIDE, PeptideAtlas, MassIVE, jPOST, iProX和Panorama将元数据的标准化XML文件提交到PX, PX提供了统一的接口来访问搜索这些元数据, 并提供原始数据的链接, 用户通过这些链接可以下载原始实验数据. 另一方面, 随着新仪器、新样品制备技术、新数据分析方法的出现, 新的数据形式不断涌现. 这一点对各个数据库提出了更高的要求, 需要各个数据库与时俱进, 实时更新. PX联盟成员定时进行数据库的升级改造, 从而适应蛋白质组技术上的更新以及724小时稳定高效的运行服务, 例如PRIDE, iProX等.

蛋白质组学资源的元数据要求目前还不如其他更

成熟的组学领域全面，导致蛋白质组学数据注释不完善的问题更加明显. 数据库维护方需要在所需的元数据数量与研究人员共享数据的意愿之间取得平衡. 由于数据共享文化在蛋白质组学中比在其他学科中起步更晚, 因此从技术和时间效率的角度来看, 迄今为止的重点是尽可能促进数据共享过程. 在这种情况下, 开发一些可以从数据文件中自动提取元数据的流程是很有必要的，在减少提交者手动输入同时可以使数据的注释信息更加完善. 为了吸引更多的研究人员提交数据, 还需要简化数据提交的界面，并对提交流程进行标准化.

数据的共享是开放数据库建设的初衷. 数据的开源共享对于生物医学领域的研究有重大影响，支持研究的经费机构以及学术杂志都大力鼓励学术研究产生数据的共享. 但是, 与实际数据产出量相比较, 实现开源共享的数据比例仍然不高，特别是在蛋白质组研究领域. 虽然目前仍然存在诸多困难和挑战，但蛋白质组学研究领域已经就数据的发布和共享达成共识，随着相应政策的不断完善、数据存储平台和数据标准的不断发展，高质量的实验原始数据和分析结果的发布和共享将形成常态化、标准化，而这毋庸置疑同时也将推动蛋白质组研究的蓬勃发展.

参考文献

1 Deutsch E W, Bandeira N, Sharma V, et al. The ProteomeXchange consortium in 2020: enabling ‘big data’ approaches in proteomics. Nucleic Acids Res, 2020, 48: D1145–D1152

2 CNCB-NGDC Members and Partners . Database resources of the National Genomics Data Center, China National Center for Bioinformation in 2021. Nucleic Acids Res, 2021, 49: D18–D28

3 Rodriguez H, Snyder M, Uhlén M, et al. Recommendations from the 2008 International Summit on proteomics data release and sharing policy: the Amsterdam Principles. J Proteome Res, 2009, 8: 3689–3692

4 Kinsinger C R, Apffel J, Baker M, et al. Recommendations for mass spectrometry data quality metrics for open access data (corollary to the Amsterdam principles). Proteomics, 2012, 12: 11–20

5 Kinsinger C R, Apffel J, Baker M, et al. Recommendations for mass spectrometry data quality metrics for open access data (corollary to the Amsterdam Principles). Mol Cell Proteomics, 2011, 10: O111.015446

6 Mayer G, Montecchi-Palazzi L, Ovelleiro D, et al. The HUPO proteomics standards initiative- mass spectrometry controlled vocabulary. Database, 2013, 2013(0): bat009

7 Mayer G, Jones A R, Binz P A, et al. Controlled vocabularies and ontologies in proteomics: overview, principles and practice. Biochim Biophys Acta, 2014, 1844: 98–107

8 Martínez-Bartolomé S, Binz P A, Albar J P. Plant Proteomics. Totowa: Humana Press, 2014. 765–780

9 Deutsch E W, Orchard S, Binz P A, et al. Proteomics standards initiative: fifteen years of progress and future work. J Proteome Res, 2017, 16: 4288–4298

10 Martens L, Chambers M, Sturm M, et al. mzML—a community standard for mass spectrometry data. Mol Cell Proteomics, 2011, 10: R110.000133

11 Vizcaíno J A, Mayer G, Perkins S, et al. The mzIdentML data standard version 1.2, supporting advances in proteome informatics. Mol Cell Proteomics, 2017, 16: 1275–1285

12 Walzer M, Qi D, Mayer G, et al. The mzQuantML data standard for mass spectrometry-based quantitative studies in proteomics. Mol Cell Proteomics, 2013, 12: 2332–2340

13 Griss J, Jones A R, Sachsenberg T, et al. The mzTab data exchange format: communicating mass-spectrometry-based proteomics and metabolomics experimental results to a wider audience. Mol Cell Proteomics, 2014, 13: 2765–2775

14 Xu Q, Griss J, Wang R, et al. jmzTab: a Java interface to the mzTab data standard. Proteomics, 2014, 14: 1328–1332

15 Perez-Riverol Y, Uszkoreit J, Sanchez A, et al. ms-data-core-api: an open-source, metadata-oriented library for computational proteomics. Bioinformatics, 2015, 31: 2903–2905

16 Qi D, Zhang H, Fan J, et al. The mzqLibrary—an open source Java library supporting the HUPO-PSI quantitative proteomics standard. Proteomics, 2015, 15: 3152–3162

17 Côté R G, Griss J, Dianes J A, et al. The PRoteomics IDEntification (PRIDE) Converter 2 Framework: an improved suite of tools to facilitate data submission to the PRIDE database and the ProteomeXchange Consortium. Mol Cell Proteomics, 2012, 11: 1682–1689

18 Desiere F. The PeptideAtlas project. Nucleic Acids Res, 2006, 34: D655–D658

19 Deutsch E W, Bandeira N, Perez-Riverol Y, et al. The ProteomeXchange consortium at 10 years: 2023 update. Nucleic Acids Res, 2023, 51: D1539–D1548

20 Deutsch E W, Csordas A, Sun Z, et al. The ProteomeXchange consortium in 2017: supporting the cultural change in proteomics public data deposition. Nucleic Acids Res, 2017, 45: D1100–D1106

21 Perez-Riverol Y, Csordas A, Bai J, et al. The PRIDE database and related tools and resources in 2019: improving support for quantification data. Nucleic Acids Res, 2019, 47: D442–D450

22 Deutsch E W, Lam H, Aebersold R. PeptideAtlas: a resource for target selection for emerging targeted proteomics workflows. EMBO Rep, 2008, 9: 429–434

23 Watanabe Y, Yoshizawa A C, Ishihama Y, et al. The jPOST Repository as a public data repository for shotgun proteomics. In: Carrera M, Mateos J. eds. Shotgun Proteomics. Methods in Molecular Biology. New York: Humana Press, 2021. 309–322

24 Moriya Y, Kawano S, Okuda S, et al. The jPOST environment: an integrated proteomics data repository and database. Nucleic Acids Res, 2019, 47: D1218–D1224

25 Okuda S, Watanabe Y, Moriya Y, et al. jPOSTrepo: an international standard data repository for proteomes. Nucleic Acids Res, 2017, 45: D1107–D1111

26 Chen T, Ma J, Liu Y, et al. iProX in 2021: connecting proteomics data sharing with big data. Nucleic Acids Res, 2022, 50: D1522–D1527

27 Ma J, Chen T, Wu S, et al. iProX: an integrated proteome resource. Nucleic Acids Res, 2019, 47: D1211–D1217

28 Sharma V, Eckels J, Schilling B, et al. Panorama public: a public repository for quantitative data sets processed in skyline. Mol Cell Proteomics, 2018, 17: 1239–1244

29 Kirkwood K I, Pratt B S, Shulman N, et al. Utilizing Skyline to analyze lipidomics data containing liquid chromatography, ion mobility spectrometry and mass spectrometry dimensions. Nat Protoc, 2022, 17: 2415–2430

30 Smith B E, Hill J A, Gjukich M A, et al. Data Mining In Proteomics. Totowa: Humana Press, 2011. 123–145

31 Craig R, Cortens J P, Beavis R C. Open source system for analyzing, validating, and storing protein identification data. J Proteome Res, 2004, 3: 1234–1242

32 Wen B, Zhang B. PepQuery2 democratizes public MS proteomics data for rapid peptide searching. Nat Commun, 2023, 14: 2213

33 Wang M, Weiss M, Simonovic M, et al. PaxDb, a database of protein abundance averages across all three domains of life. Mol Cell Proteomics, 2012, 11: 492–500

34 Perez-Riverol Y, Alpi E, Wang R, et al. Making proteomics data accessible and reusable: current state of proteomics databases and repositories. Proteomics, 2015, 15: 930–950

35 Nanjappa V, Thomas J K, Marimuthu A, et al. Plasma Proteome Database as a resource for proteomics research: 2014 update. Nucl Acids Res, 2014, 42: D959–D965

36 Sun Q, Zybailov B, Majeran W, et al. PPDB, the Plant Proteomics Database at Cornell. Nucleic Acids Res, 2009, 37: D969–D974

37 Komatsu S. Rice Proteome Database: a step toward functional analysis of the rice genome. Plant Mol Biol, 2005, 59: 179–190

38 Muthye V, Kandoi G, Lavrov D V. MMPdb and MitoPredictor: tools for facilitating comparative analysis of animal mitochondrial proteomes. Mitochondrion, 2020, 51: 118–125

39 Shifman M A, Li Y, Colangelo C M, et al. YPED: a web-accessible database system for protein expression analysis. J Proteome Res, 2007, 6: 4019–4024

40 Dinkel H, Chica C, Via A, et al. Phospho.ELM: a database of phosphorylation sites—update 2011. Nucleic Acids Res, 2011, 39: D261–D267

41 Diella F, Gould C M, Chica C, et al. Phospho.ELM: a database of phosphorylation sites update 2008. Nucleic Acids Res, 2007, 36: D240–D244

42 Diella F, Cameron S, Gemünd C, et al. Phospho.ELM: A database of experimentally verified phosphorylation sites in eukaryotic proteins. BMC BioInf, 2004, 5: 79

43 Hornbeck P V, Zhang B, Murray B, et al. PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. Nucleic Acids Res, 2015, 43: D512–D520

44 Gnad F, Gunawardena J, Mann M. PHOSIDA 2011: the posttranslational modification database. Nucleic Acids Res, 2011, 39: D253–D260

45 Vaudel M, Verheggen K, Csordas A, et al. Exploring the potential of public proteomics data. Proteomics, 2016, 16: 214–225

46 Martens L, Vizcaíno J A. A golden age for working with public proteomics data. Trends Biochem Sci, 2017, 42: 333–341
47 Zahn-Zabal M, Michel P A, Gateau A, et al. The neXtProt knowledgebase in 2020: data, tools and usability improvements. Nucleic Acids Res, 2019, 48: D328
48 Gillet L C, Navarro P, Tate S, et al. Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. Mol Cell Proteomics, 2012, 11: O111.016717
49 Deutsch E W, Perez-Riverol Y, Carver J, et al. Universal Spectrum Identifier for mass spectra. Nat Methods, 2021, 18: 768–770
50 Lund-Johansen F, de la Rosa Carrillo D, Mehta A, et al. MetaMass, a tool for meta-analysis of subcellular proteomics data. Nat Methods, 2016, 13: 837–840
51 Dong H, Liu Y, Zeng W, et al. A deep learning-based tumor classifier directly using MS raw data. Proteomics, 2020, 20: 1900344
52 Jiang Y, Sun A, Zhao Y, et al. Proteomics identifies new therapeutic targets of early-stage hepatocellular carcinoma. Nature, 2019, 567: 257–261
53 Matic I, Ahel I, Hay R T. Reanalysis of phosphoproteomics data uncovers ADP-ribosylation sites. Nat Methods, 2012, 9: 771–772
54 Hahne H, Kuster B. Discovery of O-GlcNAc-6-phosphate modified proteins in large-scale phosphoproteomics data. Mol Cell Proteomics, 2012, 11: 1063–1069

Current status and applications of proteomic databases

LIU Yi, XU XiaoFang, MA Jie, CHEN Tao & ZHU YunPing

State Key Laboratory of Proteomics, Beijing Proteome Research Center, National Center for Protein Sciences (Beijing), Beijing Institute of Lifeomics, Beijing 102206, China

The development and application of proteomic technology make the need for proteome data sharing increasingly urgent. Supported by the concept of data sharing, proteomic databases continue to emerge. The databases effectively maintain the consistency and integrity of the data information, reduce data redundancy, and facilitate effective retrieval and access by users. This paper first introduces the status of proteomic databases, focuses on the International Proteomic Data Consortium and its members, introduces the Chinese proteomic database iProX in detail, then briefly introduces the application of proteomic database, and finally looks forward to the challenges and opportunities in this field.

proteomics, database, data sharing

doi: 10.1360/SSV-2023-0141