ICS 07.080
CCS C 07

团　　体　　标　　准

T/SZAS 40—2021

蛋白质组学数据集

Dataset of proteomics

2021-09-27 发布　　2021-09-28 实施

深圳市标准化协会　发布

目　次

前　言

本文件按照GB/T 1.1—2020《标准化工作导则　第1部分：标准化文件的结构和起草规则》的规定起草。

请注意本文件的某些内容可能涉及专利。本文件的发布机构不承担识别专利的责任。

本文件由深圳市标准化协会提出并归口。

本文件起草单位：深圳华大生命科学研究院、深圳华大基因股份有限公司、广州中医药大学、西北农林科技大学、青岛华大基因研究院、深圳市坪山区尼奥基因组学研究院、深圳市易基因科技有限公司、深圳华大基因科技有限公司。

本文件主要起草人：游丽金、陈凤珍、戚达、杨晓萍、郑夏生、姜雨、郭学芹、魏晓锋、张敏文、华聪、刘姗姗、胡琪、王君文、曾文君、李良、李启沅、王博、王韧、吴昊、李倩一。

蛋白质组学数据集

1 范围

本文件规定了蛋白质组学数据范围、数据元的规范化定义及数据元目录、归档目录和数据格式要求。

本文件适用于蛋白质组学数据集信息的存储、治理、交换与共享。

2 规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中，注日期的引用文件，仅该日期对应的版本适用于本文件；不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

GB/T 18391.1 信息技术 元数据注册系统(MDR) 第1部分：框架

3 术语与定义

以下术语和定义适用于本文件。

3.1

数据元 Data Element

由一组属性规定其定义、标识、表示和允许值的数据单元。

[来源：GB/T 18391.1]

3.2

归档 Archive

是指将处理完并且具有保存价值的数据信息或数据文件提交到系统进行保存备案的过程，可供文章发表引用、公开检索。

4 缩略语

下列缩略语适用于本文件。

S：字串型（string）

DT：日期时间型（datetime）

5 数据元目录

5.1 数据元目录公用属性

数据元目录公用属性应符合表1的要求。

表1　数据元目录公用属性

属性名称	描述
中文姓名	注册用户中文姓名
英文姓名	注册用户英文姓名
邮箱	注册用户邮箱
手机	注册用户手机号码
部门	注册用户所在部门
组织/单位名称	公司/组织机构名称
街道	公司/组织机构所在街道
城市	公司/组织机构所在城市
国家/地区	公司/组织机构所在国家/地区

5.2　数据元目录专用属性

5.2.1　蛋白质组学数据元目录专用属性应包括项目信息、样本实验信息和质谱检测分析信息，具体数据元目录参考附录 A。部分数据元允许值宜以数据元值域代码形式表示，参考附录 B。

5.2.2　项目信息应为描述项目的数据元，如项目标题、公开日期等。

5.2.3　样本实验信息应为描述样本的数据元，如细胞类型、疾病、组织等。

5.2.4　质谱检测分析信息应为描述质谱检测和分析过程中的数据元，如数据库搜索软件名称、数据库搜索软件版本、蛋白质定量软件名称、蛋白质定量软件版本等。

6　数据归档目录

6.1　归档目录专用属性

蛋白质组学归档目录专用属性应符合表2的要求。

表2　蛋白质组学归档目录专用属性

国际标识符	英文名称	中文名称	信息保护	数据类型	数据元允许值示例
DE07.01.001.00	title	标题	不保护	S	Missing protein discovery in the cell line of D283 Med
DE07.01.002.00	description	描述	不保护	S	Missing protein discovery in the cell line of D283 Med derived from a peritoneal metastatic medulloblastom
DE07.01.004.00	announcedate	公告日期	不保护	DT	2020-09-07
DE07.01.005.00	announcement XML	公告	不保护	S	CNPro0001001000.xml
DE07.01.006.00	digital object identifier	数字标识符	不保护	S	10.26036/CNP0001285
DE07.01.010.00	Primary submitter	主要提交者	不保护	S	meimei han
DE07.01.011.00	Species list	物种列表	不保护	S	homo sapiens
DE07.01.012.00	Data type	数据类型	不保护	S	Proteome

6.2　原始数据归档目录结构

蛋白质组学原始数据归档目录结构应符合表3的要求。

表3　蛋白质组学原始数据归档目录结构

第一级	第二级	第三级	第四级
project_id	sample_id	experiment_id	run_id
项目编号	样本编号	实验编号	质谱上机检测编号

7　数据格式标准

7.1　原始数据文件

蛋白质组上传的数据应包括原始数据文件。原始数据文件的数据格式标准应符合表4的要求。

表4　原始数据文件数据格式标准

数据格式类型	数据格式名称	数据格式标识符	数据格式允许值	示例
原始数据格式	RAW	DF04.01.001	SF04.01.001 RAW	-
	WIFF	DF04.01.001	SF04.02.002 WIFF	-
	MZML	DF04.02.001	SF04.02.001 MZML	参考附录 C
	MZXML	DF04.02.002	SF04.02.002 MZXML	-
	MGF	DF04.02.003	SF04.02.003 MGF	参考附录 D

7.2　鉴定结果文件、定量结果文件和蛋白数据库文件

蛋白质组上传的数据宜包括鉴定结果文件、定量结果文件和蛋白质数据库文件。鉴定结果文件、定量结果文件和蛋白质数据库文件宜采用表5的数据格式标准。

表5　鉴定结果文件、定量结果文件和蛋白质数据库文件数据格式标准

数据格式类型	数据格式名称	数据格式标识符	数据格式允许值	示例
鉴定结果文件格式	MZIDENTML	DF04.03.001	SF04.03.001 MZIDENTML	参考附录 E
	CSV	DF04.03.002	SF04.03.002 CSV	-
	TXT	DF04.03.003	SF04.03.003 TXT	-
	DAT	DF04.03.004	SF04.03.004 DAT	-
定量结果文件格式	MZTAB	DF04.04.001	SF04.04.001 MZTAB	-
	MZQUANTML	DF04.04.002	SF04.04.002 MZQUANTML	-
蛋白质数据库文件格式	FASTA	DF04.02.001	SF04.02.001 FASTA	参考附录 F
	PEFF	DF04.02.002	SF04.02.002 PEFF	-
注：根据实验目的和实验方法的不同，部分文件是可选的。 示例 1：只做鉴定的实验，不需要定量文件。 示例 2：采用 *de novo* 鉴定方法的，不需要蛋白质数据库文件。				

附 录 A
（资料性）
数据元目录

A.1 简介

本附录说明了推荐性数据元的标识符，名称，定义，信息保护，单位，数据类型和数据元允许值。当有新的数据元加入时可以顺延排入。

A.2 项目信息

项目信息如表A.1所示。

表A.1 项目信息

标识符	名称	定义	信息保护	单位	数据类型	数据元允许值
P06.01.001.00	项目标题	项目的标题。	不保护	-	S	-
P06.01.002.00	公开描述	概括描述项目的信息。	不保护	-	S	-
P06.01.003.00	公开日期	项目的公开日期。	不保护	-	DT	-

A.3 样本实验信息

样本实验信息如表A.2所示。

表A.2 样本实验信息

标识符	名称	定义	信息保护	单位	数据类型	数据元允许值
DE06.01.001.00	实验标题	实验的标题。	不保护	-	S	-
DE06.01.002.00	物种	物种的分类编号（Taxonomy id）或物种名称。	不保护	-	S	-
DE06.01.003.00	组织	组织的名称。	不保护	-	S	-
DE06.01.004.00	细胞类型	细胞的类型。	不保护	-	S	-
DE06.01.005.00	疾病	疾病的名称。	不保护	-	S	-
DE06.01.006.00	定量方法	蛋白质定量的方法。	不保护	-	S	B.1 定量方法的名称代码
DE06.01.007.00	酶解方法	酶解的方法。	不保护	-	S	-
DE06.01.008.00	蛋白质翻译后修饰	蛋白质翻译后修饰的方式。	不保护	-	S	-
DE06.01.009.00	实验方式	实验的方式。	不保护	-	S	B.4 实验方式的名称代码
DE06.01.010.00	MS 仪器	质谱仪器的名称。	不保护	-	S	-
DE06.01.011.00	实验过程描述	详细的实验过程描述信息。	不保护	-	S	-
DE06.01.012.00	其他信息	补充描述不含在实验过程信息中的其他信息，包括但不限于酶解、分级、富集（针对翻译后修饰项目）、质谱、数据分享等过程。	不保护	-	S	-
DE06.01.013.00	实验总结	对实验检测结果的总结，包括但不限于鉴定到的谱图数，肽段数，蛋白质数以及定量到的蛋白质等信息。	不保护	-	S	-
DE06.01.014.00	元信息公开时间	蛋白质元数据的公开时间。	不保护	-	DT	-

A.4 质谱检测分析信息

质谱检测分析信息如表A.3所示。

表A.3 质谱检测分析

标识符	名称	定义	信息保护	单位	数据类型	数据元允许值
DE08.01.001.00	数据库搜索软件名称	蛋白质组学数据库检索软件的名称。	不保护	-	S	-
DE08.01.002.00	数据库搜索软件版本	蛋白质组学数据库检索软件的版本。	不保护	-	S	-
DE08.01.003.00	搜索的数据库类型	蛋白质组学搜索的数据库类型。	不保护	-	S	-
DE08.01.004.00	搜索的数据库版本	蛋白质组学搜索的数据库版本。	不保护	-	S	-
DE08.01.005.00	诱饵数据库类型	诱饵数据库的类型。	不保护	-	S	-
DE08.01.006.00	肽段鉴定软件名称	信息分析过程中肽段鉴定软件名称。	不保护	-	S	B.3 肽段鉴定软件的名称代码
DE08.01.007.00	肽段鉴定软件版本	信息分析过程中肽段鉴定软件版本号。	不保护	-	S	-
DE08.01.008.00	肽段鉴定软件参数	信息分析过程中肽段鉴定软件参数信息。	不保护	-	S	-
DE08.01.009.00	蛋白质定量软件名称	信息分析过程中蛋白质定量软件名称。	不保护	-	S	B.2 蛋白质相对定量分析软件的名称代码
DE08.01.010.00	蛋白质定量软件版本	信息分析过程中蛋白质定量软件版本号。	不保护	-	S	-
DE08.01.011.00	蛋白质定量软件参数	信息分析过程中蛋白质定量软件参数信息。	不保护	-	S	-

附 录 B
（资料性）
数据元值域代码表

B.1 定量方法的名称代码

定量方法名称代码规定了定量方法名称的代码。
采用2位数字顺序代码，从“00”开始编码，按升序排列，见表B.1。

表B.1 定量方法名称代码表

代码	定量方法	说明
00	iTRAQ	全称isobaric Tags for Relative and Absolute Quantitation。采用4种或8种同位素的标签，通过特异性标记多肽的氨基基团，接着进行串联质谱分析，可同时比较4种或8种不同样品中蛋白质的相对含量或绝对含量。
01	Label_free	不需要对比较样本做特定标记处理，只需要比较特定肽段或蛋白质在不同样品间的色谱质谱响应信号便可得到样品间蛋白质表达量的变化，通常用于分析大规模蛋白质鉴定和定量时所产生的质谱数据。
02	MRM	一种基于已知信息或假定信息有针对性地获取数据，进行质谱信号采集的技术。
03	SILAC	细胞培养条件下稳定同位素标记技术，全称Stable Isotope Labeling by Amino Acids in Cell Culture。在细胞培养基中加入轻、中或重型稳定同位素标记的必需氨基酸(赖氨酸和精氨酸)，通过细胞的正常代谢，使新合成的蛋白质带上稳定同位素标签；等量混合各类型蛋白质,酶解后进行质谱分析；通过比较一级质谱图中同位素峰型的面积大小进行相对定量，同时利用二级谱图对肽段进行序列测定从而进行蛋白质鉴定的方法。
04	SWATH	该技术将整个质谱扫描质量范围分为若干小窗口，依次对每个窗口的所有离子进行碎裂，使其能够对扫描区间内的所有肽段离子进行高速一级MS扫描再进行二级MS/MS分析。
05	TMT	目前差异蛋白质定量分析方法中通量最高、系统误差最小、功能最强大的分析方法之一。可以做到16个样品（16-plex）同时标记分析，可以同时比较2-10组样品的蛋白质表达差异，能够提供更精确的数字化信号，更高的检测通量以及更广泛的检测范围。
06	Other	其他定量方式。

B.2 蛋白质相对定量分析软件的名称代码

蛋白质相对定量分析软件名称代码规定了蛋白质相对定量分析软件名称的代码。
采用2位数字顺序代码，从“00”开始编码，按升序排列，见表B.2。

表B.2 蛋白质相对定量分析软件名称代码表

代码	蛋白质相对定量分析软件	参考链接
00	aLFQ	https://cran.rstudio.com/web/packages/aLFQ/index.html
01	APEX	https://sourceforge.net/projects/apexqpt/
02	ASAPRatio	http://tools.proteomecenter.org/wiki/index.php?title=Software:ASAPRatio
03	Census	http://fields.scripps.edu/yates/wp/?page_id=824
04	DanteR	https://omics.pnl.gov/software/danter
05	GProX	http://gprox.sourceforge.net/
06	ICPL_ESIQuant	https://sourceforge.net/projects/icplquant/files/
07	iQuant	https://sourceforge.net/projects/iquant/
08	isobar	https://github.com/fbreitwieser/isobar
09	IsobariQ	https://pubmed.ncbi.nlm.nih.gov/21067241/
10	IsoQuant	https://pubmed.ncbi.nlm.nih.gov/22519468/
11	Libra	http://tools.proteomecenter.org/wiki/index.php?title=Software:Libra
12	MapQuant	https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/10.1002/pmic.200500201
13	Mascot	https://www.matrixscience.com/
14	MassChroQ	http://pappso.inrae.fr/bioinfo/masschroq/
15	MaxQuant	https://www.maxquant.org/
16	MFPaQ	http://mfpaq.sourceforge.net/
17	MSQuant	http://msquant.sourceforge.net/
18	MS-Spectre	https://sourceforge.net/projects/ms-spectre/
19	Multi-Q	http://ms.iis.sinica.edu.tw/COmics/Software_Multi-Q2.html
20	Multi-Q Web Server	http://ms.iis.sinica.edu.tw/COmics/Software_Multi-Q2.html
21	openMS	https://www.openms.de/
22	PEAKS-Q	https://www.bioinfor.com/quantification/
23	ProRata	https://code.google.com/archive/p/prorata/downloads
24	PVIEW	https://compbio.cs.princeton.edu/pview/
25	Quant	https://sourceforge.net/projects/protms/files/Quant/
26	XPRESS	http://tools.proteomecenter.org/wiki/index.php?title=Software:XPRESS
27	MALDIquant	http://www.strimmerlab.org/software/maldiquant/
28	ProteoIQ	http://www.premierbiosoft.com/protein_quantification_software/index.html
29	Other	-

B.3 肽段鉴定软件的名称代码

肽段鉴定软件名称代码规定了肽段鉴定软件名称的代码。
采用2位数字顺序代码，从“00”开始编码，按升序排列，见表B.3。

表B.3　肽段鉴定软件名称代码表

代码	肽段鉴定软件	参考链接
00	CPFP	https://cpfp.sourceforge.io/
01	IPeak	https://github.com/fghali/mzidlib/blob/master/IPeak%20v1.0.1.zip
02	Mascot	https://www.matrixscience.com/
03	MASSHUNTER	https://www.agilent.com.cn/zh-cn/product/software-informatics/mass-spectrometry-software
04	MS-GF+	https://github.com/MSGFPlus/msgfplus
05	Myrimatch	https://lab.vanderbilt.edu/msrc-bioinformatics/myrimatch-source/
06	Omssa	https://proteomicsresource.washington.edu/protocols06/omssa.php
07	PEAKS	https://www.bioinfor.com/peaks-online/
08	Percolator	http://percolator.ms/
09	pFind	http://pfind.ict.ac.cn/
10	ProteoCloud	https://code.google.com/archive/p/proteocloud/
11	Proteome Discoverer software	https://www.thermofisher.cn/order/catalog/product/OPTON-30810#/OPTON-30810
12	TOPP	https://www.openms.de/proteomics/
13	TPP	http://ww.peptideatlas.org/
14	X!Tandem	https://www.thegpm.org/tandem/
15	Other	-

B.4　实验方式的名称代码

实验方式名称代码规定了实验方式名称的代码。

采用2位数字顺序代码，从“00”开始编码，按升序排列，见表B.4。

表B.4　实验方式名称代码表

代码	实验方式名称
00	Top-down proteomics
01	Shotgun proteomics
02	Gel-based proteomics
03	Cross-link (CX-MX)
04	Affinity purification (AP-MS)
05	SRM/MRM
06	SWATH MS (Data-independent acquisition)
07	MSE (Data-independent acquisition)
08	HDMSE (Data-independent acquisition)
09	PAcIFIC (Data-independent acquisition)
10	All-ion fragmentation (Data-independent acquisition)
11	MS imaging
12	Other

附 录 C
（资料性）
蛋白质组质谱数据 mzML 文件

C.1 简介

mzML 格式是由人类蛋白质组组织下属的蛋白质组学标准倡议小组（HUPO-PSI, http://psidev.info）结合.mzXML和.mzData两种格式，提出的新一代蛋白质组学原始数据开源标准格式。mzML采用可扩展标记语言（eXtensible Markup Language，XML）存储数据，采用XML框架定义语言（XML Schema Definition，XSD）自定义多个元素标签及其结构。

C.2 数据格式

文件内容由<mzML></mzML>标签作为开头和结尾，中间由自定义的标签嵌套包含对应的元数据，信号数据等。其主要标签有<cvList>（受控词表），<sampleList>（样本列表），<run>（单次连续质谱扫描）。

C.3 访问地址

——mzML 文件文档地址：https://www.psidev.info/mzML；
——mzML 完整文件访问地址：
http://db.systemsbiology.net/projects/PSI/mzML/tiny1.mzXML2.0.mzXML。

附 录 D
（资料性）
蛋白质谱图 MGF 文件

D.1 简介

MGF文件为文本格式，目的是提供标准的质谱原始数据。与mzML相比，MGF文件更为简洁，只提供二级谱图的信息，也省略了很多元数据。

D.2 数据格式

MGF文件包含多个二级谱图信息。每个谱图始于‘BEGIN IONS’行，终于‘END IONS’行。包含TITLE行、PEPMASS行、CHARGE行、RTINSECONDS行、SCANS行，分别记录谱图的一般信息、母离子质量数、母离子电荷数、保留时间（秒）、扫描序号。数据多行记录，每一行由两列数字构成，第一列数字记录离子的质量数，第二列数字记录离子的信号强度。

D.3 示例

```
BEGIN IONS
TITLE=File127389 Spectrum2 scans: 3
PEPMASS=599.32275 20686.39648
CHARGE=3+
RTINSECONDS=0
SCANS=3
131.13695 2815.74
219.82968 6798.9
321.23926 5523.5
1766.84021 4106.19
END IONS
```

附 录 E
（资料性）
蛋白质鉴定 mzIdentML 文件

E.1 简介

mzIdentML文件为文本格式，后缀名为mzid，目的是用于存储记录蛋白质鉴定结果信息。本格式由人类蛋白质组组织下属的蛋白质组学标准倡议小组（HUPO-PSI, http://psidev.info）推出，成为目前公认的蛋白质组学鉴定结果的标准格式。mzIdentML采用可扩展标记语言（eXtensible Markup Language，XML）存储数据，采用XML框架定义语言（XML Schema Definition，XSD）自定义多个元素标签及其结构。

E.2 数据格式

文件内容由<MzIdentML></MzIdentML>标签作为开头和结尾，中间由自定义的标签嵌套包含对应的元数据，谱图数据，肽段数据，统计结果等。其主要标签有<cvList>（受控词表），<SequenceCollection>（蛋白质序列集合），<DataCollection>（输入输出数据集合）。

E.3 完整文件访问地址

mzIdentML完整文件访问地址：https://www.psidev.info/mzidentml。

附 录 F
（资料性）
蛋白质数据库 FASTA 文件

F.1 简介

FASTA文件为文本格式，目的是用于存储记录蛋白质序列信息。本格式包含一行蛋白质ID及注释信息，以及一行序列信息。

F.2 数据格式

序列文件的第一行以大于号‘>’开头，紧跟序列ID，以及相关的元数据（注释，基因，物种，数据库来源等）。第二行为蛋白完整氨基酸序列字符串，每一个字母代表一个氨基酸。

F.3 示例

示例1：Uniprot 数据库

```
>sp|Q5TM83|NANOG_MUSMM Homeobox protein NANOG OS=Mus musculus molossinus OX=57486 GN=Nanog PE=2 SV=1
MSVGLPGPHSLPSSEEASNSGNASSMPAVFHPENYSCLQGSATEMLCTEAASPRPSSEDLPLQGSPDSSTSPKQKLSSPEADKGPEE
EENKVLARKQKMRTVFSQAQLCALKDRFQKQKYLSLQQMQELSSILNLSYKQVKTWFQNQRMKCKRWQKNQWLKTSNGLIQKGSAPV
EYPSIHCSYPQGYLVNASGSLSMWGSQTWTNPTWSSQTWTNPTWNNQTWTNPTWSSQAWTAQSWNGQPWNAAPLHNFGEDFLQPYIQ
LQQNSSASDLEVNLEATRESHAHFSTPQALELFLNYSVTPPGEI
```

示例2：NCBI RefSeq 数据库

```
>NP_001128684.1 peptidyl-prolyl cis-trans isomerase FKBP7 isoform b precursor [Homo sapiens]
MPKTMHFLFRFIVFFYLWGLFTAQRQKKEESTEEVKIEVLHRPENCSKTSKKGDLLNAHYDGYLAKDGSKFYCSRTQNEGHPKWFVL
GVGQVIKGLDIAMTDMCPGEKRKVVIPPSFAYGKEGYEGKIPPDATLIFEIELYAVTKGPRSIETFKQIDMDNDRQLSKAEINLYLQ
REFEKDEKPRDKSYQDAVLEDIFKKNDHDGDGFISPKEYNVYQHDEL
```

参 考 文 献

[1] WS/T 306 卫生信息数据集分类与编码规则

[2] WS 363.1 卫生信息数据元目录 第1部分：总则

[3] Ma J, Chen T, Wu S, Yang C, Bai M, Shu K, Li K, Zhang G, Jin Z, He F, Hermjakob H,Zhu Y. (2019) iProX: an integrated proteome resource. Nucleic Acids Res, 47, D1211-D1217.

[4] Stephen K Burley, Charmi Bhikadiya,et al. RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. Nucleic Acids Research, 49, D437–D451.

[5] The UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Research, 49, D480–D489.

[6] Jones, A., Eisenacher, M., Mayer, G., Kohlbacher, O., Siepen, J., Hubbard, S., Selley, J., Searle, B., Shofstahl, J., Seymour, S., Julian, R., Binz, P., Deutsch, E., Hermjakob, H., Reisinger, F., Griss, J., Vizcaíno, J., Chambers, M., Pizarro, A. and Creasy, D., 2012. The mzIdentML Data Standard for Mass Spectrometry-Based Proteomics Results. Molecular & Cellular Proteomics, 11(7), pp.M111.014381-1-M111.014381-10.

[7] mzData. http://psidev.info/index.php?q=node/80#mzdata.