生 物 工 程 学 报
Chinese Journal of Biotechnology
http://journals.im.ac.cn/cjbcn
DOI: 10.13345/j.cjb.230079

Sep. 25, 2023, 39(9): 3579-3593

·综 述·

基于数据非依赖采集的以肽为中心分析算法和软件的研究进展

张莹莹[1,2]，舒坤贤[1*]，常乘[2*]

1 重庆邮电大学生物信息学院 大数据生物智能重庆市重点实验室，重庆 400065
2 北京蛋白质组研究中心 国家蛋白质科学中心(北京) 北京生命组学研究所，北京 102206

张莹莹, 舒坤贤, 常乘. 基于数据非依赖采集的以肽为中心分析算法和软件的研究进展[J]. 生物工程学报, 2023, 39(9): 3579-3593.
ZHANG Yingying, SHU Kunxian, CHANG Cheng. Advances of peptide-centric data-independent acquisition analysis algorithms and software tools[J]. Chinese Journal of Biotechnology, 2023, 39(9): 3579-3593.

摘 要：数据非依赖采集(data-independent acquisition, DIA)是一种高通量、无偏性的质谱数据采集方法，具有定量结果重现性好，对低丰度蛋白质友好的特点，是近年来进行大队列蛋白质组研究的首选方法之一。由于DIA产生的二级谱是混合谱，包含了多个肽段的碎片离子信息，使得蛋白质鉴定和定量更加困难。目前，DIA数据分析方法分为两大类，即以肽为中心和以谱图为中心。其中，以肽为中心的分析方法鉴定更灵敏，定量更准确，已成为DIA数据解析的主流方法。其分析流程包括构建谱图库、提取色谱峰群、特征打分和结果质控4个关键步骤。本文综述了以肽为中心的DIA数据分析流程，介绍了基于此流程的数据分析软件及相关比较评估工作，进一步总结了已有的算法改进工作，最后对未来发展方向进行了展望。

关键词：计算蛋白质组学；定量蛋白质组学；数据非依赖采集；以肽为中心

资助项目：国家重点研发计划(2021YFA1301603)
This work was supported by the National Key Research and Development Program of China (2021YFA1301603).
*Corresponding authors. E-mail: SHU Kunxian, shukx@cqupt.edu.cn; CHANG Cheng, changcheng@ncpsb.org.cn
Received: 2023-02-06; Accepted: 2023-04-11; Published online: 2023-04-13

Advances of peptide-centric data-independent acquisition analysis algorithms and software tools

ZHANG Yingying[1,2], SHU Kunxian[1*], CHANG Cheng[2*]

1 Chongqing Key Laboratory of Big Data for Bio-intelligence, School of Bioinformatics, Chongqing University of Posts and Telecommunications, Chongqing 400065, China
2 Beijing Proteome Research Center, National Center for Protein Science (Beijing), Beijing Institute of Life Omics, Beijing 102206, China

Abstract: Data-independent acquisition (DIA) is a high-throughput, unbiased mass spectrometry data acquisition method which has good quantitative reproducibility and is friendly to low-abundance proteins. It becomes the preferred choice for clinical proteomic studies especially for large cohort studies in recent years. The mass-spectrometry (MS)/MS spectra generated by DIA is usually heavily mixed with fragment ion information of multiple peptides, which makes the protein identification and quantification more difficult. Currently, DIA data analysis methods fall into two main categories, namely peptide-centric and spectrum-centric. The peptide-centric strategy is more sensitive for identification and more accurate for quantification. Thus, it has become the mainstream strategy for DIA data analysis, which includes four key steps: building a spectral library, extracting ion chromatogram, feature scoring and statistical quality control. This work reviews the peptide-centric DIA data analysis procedure, introduces the corresponding algorithms and software tools, and summarizes the improvements for the existing algorithms. Finally, the future development directions are discussed.

Keywords: computational proteomics; quantitative proteomics; data-independent acquisition; peptide-centric

蛋白质组的概念于 1994 年由 Marc Wilkins 提出[1]，是蛋白质(protein)和基因组(genome)的结合，是指一个基因组、细胞和组织所表达的所有蛋白质。蛋白质组学旨在对一定条件下产生的蛋白质组进行研究，包括蛋白质的差异表达、活性的改变以及翻译后修饰。只有全面了解不同组织、器官在不同生理、病理状态下蛋白质组的构成和变化，才能理解各种生命活动的机制，从而把握疾病预防和治疗的关键。在人类基因组计划草图公布的 10 年之后，人类蛋白质组组织(Human Proteome Organization, HUPO)于 2010 年启动了人类蛋白质组计划[2]。截至 2020 年，人类蛋白质组计划绘制了人体中超过 90%的高质量人类蛋白质组图谱[3]。在掌握了如此多的蛋白质密码之后，蛋白质组研究的重点在于这些蛋白质如何在生物系统中发挥作用。

目前质谱技术因其高通量、高准确性已成为蛋白质组学研究的首选，在疾病机制研究、生物标志物筛选等诸多研究中发挥了重要作用。基于质谱的自底而上的蛋白质研究策略，主要分为 3 步：样品制备、液相色谱-质谱联用技术采集数据和数据分析[4]。样品制备是指从样品中提取蛋白质，并采用序列特异性酶(如 trypsin、Glu-C 和 Lys-C 等)进行酶切，从而将

蛋白混合物酶解成肽段混合物。液相色谱-质谱联用技术通过液相色谱将不同物理化学性质的肽段分离，依次进入质谱仪，进入到质谱的肽段被电离，电离后的肽段(母离子)经质荷比检测得到一级谱图，一级谱图中的母离子经质量分析器选择并碎裂为子离子，也称碎片离子，子离子经质荷比检测获得二级谱图。

目前有 2 种常用的数据采集模式，即数据依赖采集(data-dependent acquisition, DDA)和数据非依赖采集(data-independent acquisition, DIA)。DDA 采集模式选择肽段离子中的高丰度信号峰进行二级碎裂，获得的二级谱中只包含了单个肽段母离子对应的子离子信息，因此 DDA 数据分析较为简单。但 DDA 更倾向于对高丰度离子采样，具有丰度依赖性和采样随机性。DIA 采集模式理论上可以对一级谱图中的所有母离子进行碎裂，不遗漏任何信息。其结果定量的重复性和准确性好，尤其适用于低丰度蛋白质的分析，引领了定量蛋白质组学新的发展方向，并已广泛应用于大队列临床蛋白质组学研究。

但 DIA 数据高度复杂，DIA 二级谱图其实是由多种母离子同时经过二级碎裂产生的混合谱，这使得 DIA 数据解析较为困难。目前，DIA 数据分析方法分为两大类：以谱图为中心和以肽为中心。以谱图为中心的方法是从二级谱图出发，通过解卷积等方法解析出每个母离子的子离子信号，后续通过 DDA 数据分析方法进行蛋白质的鉴定和定量。以肽为中心的方法则是预先构建谱图库，根据谱图库中肽段的已知信息靶向提取 DIA 数据中的母离子和子离子信号，鉴定灵敏度和定量准确度高，已被广泛用于大样本量研究中。本文详细介绍了以肽为中心分析流程中构建谱图库、提取色谱峰群、特征打分和统计质控 4 个关键步骤，总结了相关软件的特点及评估工作，可供蛋白质组学研究者分析 DIA 数据时参考。此外，本文概述了现有以肽为中心的 DIA 数据分析算法的改进和优化工作，并对未来发展方向作了展望和讨论。

1 DIA 原理与方法

DIA 是一种无偏质谱数据采集技术[5]。相较于 DDA 对肽段进行选择性采集，DIA 是扫描一级谱中所有的母离子并采集子离子信号。扫描方式分为两类：全离子碎裂扫描和分段扫描[6]。全离子碎裂扫描方式鸟枪法碰撞诱导解离(shotgun collision-induced dissociation, Shotgun CID)在 2003 年由 Purvine 等[7]提出，随后还出现了全离子碎裂(all ion fragmentation, AIF)[8]、高能质谱仪(mass spectrometry$^{\text{elevated energy}}$, MS^E)[9]等方法。全离子碎裂扫描的特点是选择一级谱中全部母离子进行碎裂，但由于不同母离子的丰度差异较大，受高丰度信号抑制效应等因素影响，从其二级谱中仅能检测出少量的高丰度肽段信息，适用于分析简单样品。为解决全离子碎裂扫描的不足，研究者相继提出了不同的分段扫描方法。分段扫描在一级谱划定多个母离子扫描窗口，每次仅碎裂扫描窗口中的母离子，扫描窗口个数覆盖一级谱质量范围。分段扫描方法降低了二级谱的复杂度，适用于分析复杂样本。其中，Gillet 等[10]在 2012 年提出的所有理论碎片离子质谱的顺序窗口采集(sequential windowed acquisition of all theoretical fragment ion mass spectra, SWATH-MS)是应用最广泛的分段扫描 DIA 技术之一。多路复用策略(multiplexing strategy, MSX)[11]、扫描四倍 DIA (scanning quadruple DIA, SONAR)[12]数据采集方法在此基础上调整了母离子扫描窗口的大小。以上的母离子扫描窗口大小是固定不变的。但由于谱图的高度复杂性，不同窗口所包

含的母离子数量不同。为了适应母离子的分布密度，超反应监测[13] (hyper reaction monitoring, HRM)、可变窗口 SWATH-MS[14]等指定可变窗口的数据采集方法被提出。这些采集方法的进展主要局限于二级谱水平，而用于一级谱扫描的离子采集仍然非常低效，无法应对血浆、组织等蛋白质动态范围大的样本。Meier 等[15]提出了一种新的数据采集方法，称为 BoxCar，其主要思想是在获取一级谱图时采用分段累积的方法，使得平均离子注入时间相较全扫描增加 10 倍以上。此方法采集的 DIA 数据其一级谱的动态范围更高，二级谱包含的肽段信号更多。

2 以肽为中心算法与软件

2.1 以肽为中心分析算法流程

以肽为中心的 DIA 数据分析算法旨在通过已知的肽段信息靶向提取质谱数据中母离子及子离子信号，从而实现肽段的鉴定与定量。如图 1 所示，其关键步骤包括构建谱图库、色谱峰提取、特征打分和统计质控。

2.1.1 构建谱图库

在 DIA 数据分析中，谱图库中保存的肽段先验信息[16] (peptide query parameters, PQPs)是进行肽段鉴定和定量的重要基础，主要包括了蛋白质的特异性肽段信息(4–12 条)：母离子和子离子质荷比(mass to charge ratio, *m*/*z*)、保留时间(retention time, RT)、母离子和子离子的电荷数(charge)、以及子离子的相对强度值(intensity)等。谱图库中的先验信息可以通过 3 种方式获取：DDA 实验建库、深度学习预测建库和 DIA 实验建库。

DDA 实验建库是指首先对样本进行 DDA 数据采集，然后根据 DDA 数据的鉴定结果获取肽谱匹配信息(peptide spectrum matches, PSMs)，用于构建谱图库。该方法构建的谱图库具有良好的特异性，但仅包含在 DDA 实验中鉴定到的

Precursor *m/z*	Fragment *m/z*	RT	Intensity	Peptide	Protein
670.37	755.441	6.5	7 865.3	AAAAAAALQAK	P36578
670.37	826.478	6.5	3 782.5	AAAAAAALQAK	P36578
670.37	698.427	6.5	8 922.6	AAAAAAALQAK	P36578
670.37	1 076.346	6.5	1 256.7	AAAAAAALQAK	P36578

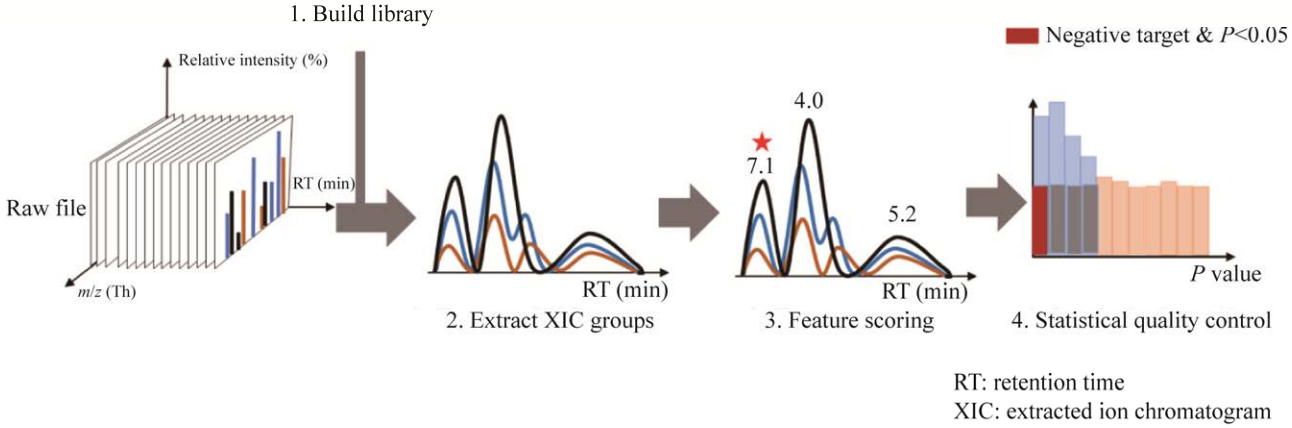

图 1 以肽为中心的 DIA 数据分析流程

Figure 1 Peptide-centric analysis pipeline of DIA.

肽段，蛋白质覆盖度有限。虽然预分离和重复实验可以部分弥补这种不足，但也会增加样品用量和质谱检测机时，因此不适用微量样本分析。此外，还可以从 SWATHAtlas 网站(http://www.swathatlas.org/)上获取基于大量不同组织样本构建的公共谱图库，例如 2014 年发表的 Pan-Human Library (PHL)[17]和 2020 年发表的 DIA Pan-Human Library (DPHL)[18]。由于在构建公共谱图库时实验参数和样本实验参数不能完全匹配，因此公共谱图库的特异性较差。

第二种生成 DIA 谱图库的方式是基于蛋白质序列信息，采用深度学习技术进行预测。2017 年 Zhou 等[19]在 *Analytical Chemistry* 上发表了 pDeep，采用循环神经网络结构，并使用了来自 3 个实验室、3 种不同碎裂方式的 8 个数据集进行模型的训练和测试。在 ProteomeTools project[20]的数据集上进行测试，结果显示预测谱图和实验谱图的平均皮尔森相关系数大于 0.9。2018 年 Ma 等[21]在 *Analytical Chemistry* 上发表了 DeepRT，采用卷积神经网络和循环神经网络的混合网络架构，使用了 4 层卷积神经网络和循环神经网络处理被离散化后的肽段序列。经过 2 种神经网络的特征提取后，利用主成分分析进行降维，最后采用 3 种机器学习方法(支持向量机、随机森林及梯度提升)预测保留时间，实现了理论预测值与真实值相关性接近 0.99。2019 年 Gessulat 等[22]在 *Nature Methods* 上发表了 Prosit，该模型由 1 个编码器和 1 个解码器组成，使用 one-hot 编码对肽段进行编码，并利用肽段表示向量进行编码和解码来学习数据的特征。该模型以肽段电荷数、归一化碎裂能量和肽段序列作为输入，以保留时间和二级谱图作为输出，利用 ProteomeTools project 项目提供的高质量谱图数据和相应的肽段鉴定结果训练和测试模型。该模型对保留时间和二级谱图的预测和实验产生的保留时间和二级谱图相关性接近 1。2020 年 Yang 等[23]在 *Nature Communications* 上发表了 DeepDIA。该模型是基于卷积神经网络和循环神经网络的深度神经网络模型，用于预测肽段的二级谱图和保留时间。该模型使用了 one-hot 编码对肽段进行编码，输出每个可能的 b、y 离子以及相对强度和肽段的保留时间。研究人员将 DeepDIA 用于未去高丰度蛋白的血清样品的 DIA 数据分析。结果显示，DeepDIA 构建的谱图库检测到蛋白质种类是 DDA 建库的 2 倍以上。2021 年 Bouwmeester 等[24]在 *Nature Methods* 上发表了 DeepLC，用于预测修饰肽段和非修饰肽段的保留时间。DeepLC 采用卷积神经网络模型，使用基于原子组成的肽编码方式。这种编码方式使模型能够学习并归纳未知的修饰肽段。在 20 个数据集上，Pearson 相关系数都能达到 0.99。尽管基于深度学习的谱图库生成软件可以实现高准确度和蛋白质覆盖率，但也会引入大量在 DIA-MS 中无法鉴定的阴性肽段，降低了谱图库的特异性。这不仅增加了 DIA 数据分析的时间，还会降低蛋白质鉴定的灵敏度[25]。

第三种建库方式是 Pino 等[26]提出的气相分馏(gas-phase fractionation, GPF)-DIA，其利用气相分馏和 DIA 技术生成的肽段-保留时间的匹配结果来构建色谱库。GPF-DIA 方案通过对相同的样品进行 6 次重复进样，并将母离子按质荷比划分为多个小区间进行质谱检测。然后利用 DIA 数据的鉴定结果来构建色谱库。相较于传统的 DDA 建库方案，GPF-DIA 可以检测到更多的目标蛋白，节约样品量和检测时间，同时提供更准确的定量结果。此外，与基于深度学习预测的谱图库相比，GPF-DIA 生成的谱图库具有更好的特异性。

2.1.2 提取色谱峰群

在谱图库构建完成后，通过谱图库中肽段

和碎片离子的 RT 和 *m*/*z* 提取色谱峰群，即肽段的离子流色谱峰(extracted ion chromatogram, XIC)和碎片离子的 XIC 组合而成的色谱峰群(图 2)。

首先，基于谱图库中提供的母离子和子离子质荷比在所有一级谱图中提取母离子的离子流色谱峰(extracted ion chromatogram, XIC)[27]，并在其母离子所在的扫描窗口内的所有二级谱图中提取子离子 XIC。考虑质谱仪稳定性有限，肽段洗脱时间会产生固有的偏差，因此需要设定保留时间范围提取 XIC。OpenSWATH[28]默认使用 RT±5 min 的提取时间窗口去提取 XIC；而 DIA-NN[29]将保留时间扩大到整个色谱时间。

构建谱图库和 DIA 实验时的色谱条件及仪器状态不可避免地存在差异，同 1 条肽段在不同实验条件下的保留时间可能不同，因此需要进行保留时间的标准化。为此，可以添加标肽[29]来实现 RT (实测保留时间)到 iRT (标准化保留时间)的转换。目前常用的标肽是人工合成的多条肽段，并且已知母离子和子离子的 *m*/*z* 以及 iRT，根据母离子和子离子 *m*/*z* 能获取每条标肽在不同实验条件下的 RT。在谱图库构建时，根据标肽的 iRT 和实验或者预测的 RT，使用最小二乘法拟合线性函数 iRT=f(RT)，通过该函数可得谱图库中所有母离子的 iRT。在使用该谱图库进行 DIA 数据分析时，根据每条标肽的 iRT 和在 DIA 实验中的 RT，通过线性拟合得到函数 RT=f(iRT)。通过该函数可得谱图库中所有母离子在该 DIA 实验中的 RT，用于提取母离子和子离子色谱峰群。以上 RT 和 iRT 相互转换的规则[30]，可以实现不同谱图库应用于不同色谱条件下的 DIA 数据。

在信号靶向提取的整个流程中，需要对肽段 RT 和 *m*/*z* 进行校正，以提高肽段鉴定的准确性和可靠性。*m*/*z* 的校正往往基于高可信肽段的实测 *m*/*z* 和理论 *m*/*z* 的质量偏差。RT 的校正可

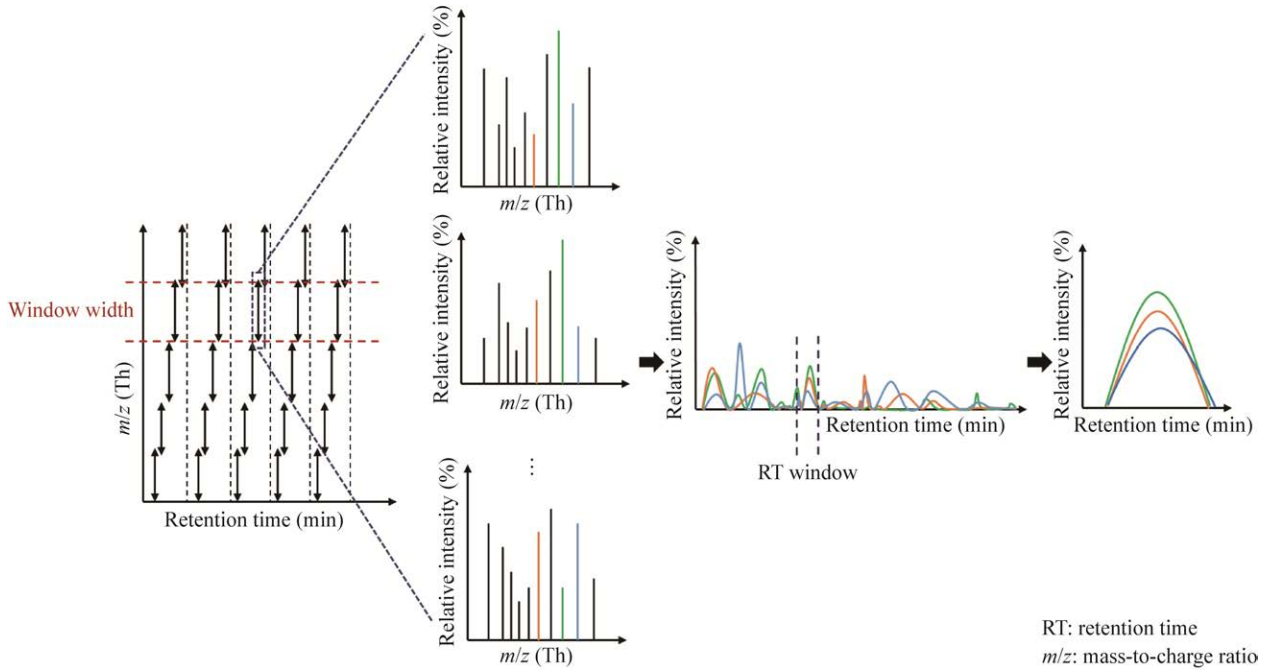

图 2　提取碎片离子的色谱峰

Figure 2　Extraction of the chromatographic peaks of the fragment ions.

利用标肽的实测 RT 和理论 RT 进行校正。以 DIA-NN 为例，其第 1 轮鉴定会在整个 RT 范围内进行，以确定每条肽段的最佳 RT。之后，只有错误发现率(false discovery rate, FDR)小于 1%的肽段被选中用于 RT 校正。

2.1.3 特征打分

提取母离子和子离子的色谱峰群后，下一步就是对其进行特征打分，用于后续评估色谱峰群对应的肽段是否正确。待计算的特征可分为以下几类：(1) 色谱层面上，可计算母离子、子离子及其同位素离子流色谱峰之间的相似性，衡量离子是否共流出。(2) 谱图库层面上，可计算谱图库中肽段的保留时间、相对强度以及 DIA 数据中实测保留时间和相对强度的一致性。同时，还需要考虑谱图库中母离子的质量数、电荷数、长度以及子离子个数等可量化的特征。(3) 谱图层面上，可计算子离子的质量准确度，b、y 离子的个数等。相似性和一致性的计算通常采用人工设计的函数，如皮尔逊相关系数和余弦距离等。不同软件计算的特征大多来自以上 3 个层面，但具体考虑的特征数目并不相同。例如，OpenSWATH[28]计算了 11 个特征，DIA-NN[29]计算了 73 个打分，MaxDIA[31]计算了 60 个特征。

2.1.4 统计质控

经过上述特征打分计算，可以得到每个色谱峰群的特征向量(x_1, x_2, x_3, ... , x_n)，接下来就需要判断该色谱峰群代表的目标肽段的正确性(y)。为此，研究人员采用了目标-诱饵肽段竞争策略[32]，通过谱图库中的目标肽段生成诱饵肽段，保证其肽段保留时间、电荷信息不变，但改变子离子的质荷比来“欺骗”鉴定软件。该策略首先生成诱饵肽段，然后将目标肽段和诱饵肽段提取的色谱峰群的特征向量作为分类器的输入，训练分类器($y=w_1x_1+w_2x_2+w_3x_3+ \dots +w_nx_n$)来区分诱饵和目标，并生成总打分($y$)。目前分类器有 2 种类型：机器学习分类器和神经网络分类器。mProphet[33]是基于机器学习方法的典型代表。它最初被用于分析靶向蛋白质组学数据，但其基本思想适用于 DIA 数据质控。OpenSWATH 使用 PyProhet[34]算法，它是对 mProphet 的 python 重实现，并对其改进，使其更适用于 DIA 数据，例如采用多个分类器结合，构建了非参数、对数正态分布的空分布。MaxDIA[31]利用集成学习，使用机器学习 XGBoost 构建分类器，它的基本思想是将弱分类器组合成强分类器，给分类效果更好的分类器更高的权重。DIA-NN 首先使用了线性判别分析(linear discriminant analysis, LDA)或者线性分类器来区分诱饵肽段色谱峰群和目标肽段色谱峰群，获取特征组合的权重，挑选肽段的最优峰，再进行深度神经网络(deep neural network, DNN)模型训练，计算 FDR[29]。

2.2 以肽为中心 DIA 数据分析软件的比较评估

随着 DIA 技术的快速发展，目前已经有不少以肽为中心的 DIA 数据分析软件应运而生。本文列举了当前常用的以肽为中心的 DIA 数据分析软件(表 1)。同时，由于不同分析软件可能会对后续蛋白质定性和定量结果产生一定的影响，本文总结了已发表的软件工具的评估和比较工作，供研究人员参考。此外，已有研究结果表明，谱图库的构建方法以及数据采集时使用的窗口大小等因素也会对蛋白质鉴定和定量结果产生影响[39-40]。

Navarro 等[41]使用了不同比例的混合物种蛋白质组数据构建了金标准数据集 LFQbench，并用该数据集评估了 4 种以肽为中心的软件 OpenSWATH、SWATH™ 2.0 (SCIEX 公司研发的 SWATH 数据分析商业软件)、Skyline、Spectronaut

表 1 以肽为中心 DIA 数据分析软件及工具列表

Table 1 List of peptide-centric software tools for DIA data analysis

Software	Year	Journal	Library	Rescore method	Feature
DreamDIA[35]	2021	*Communications Biology*	Library-based	XGBoost semi-supervised learning	Extract features using deep representation models
MaxDIA[31]	2021	*Nature Biotechnology*	Library-based Library-free	XGBoost semi-supervised learning	Perform multiple rounds of matching with Bootstrap
DIA-NN[29]	2019	*Nature Methods*	Library-based Library-free	Ensemble DNN	Select the best chromatographic peaks using linear discriminant analysis
ScaffoldDIA[36]	2018	*Nature Communication*	Library-based	Percolator	Support the use of chromatographic libraries
OpenSWATH[28]	2014	*Nature Biotechnology*	Library-based	PyProphet	Integrate multiple analysis modules
Spectronaut[37]	2014	*60th American Society for Mass Spectrometry Conference*	Library-based Library-free	PyProphet	Commercial software
Skyline[38]	2010	*Bioinformatics*	Library-based	mProphet	Visualize chromatographic peak groups

和 1 种以谱图为中心的软件 DIA-Umpire[42]，在 1% FDR 下评估了定量准确度、肽段鉴定灵敏度、鲁棒性以及特异性。评估结果表明 OpenSWATH、Spectronaut 和 Skyline 对低丰度肽段的定量值和理论值存在偏差，而 DIA-Umpire、SWATH 2.0 定量结果偏差较小。

Gotti 等[39]使用不同大小的采集窗口、不同比例混合的样品以及不同来源的谱图库，系统评估了 DIA-NN、DIA-Umpire、OpenSWATH、ScaffoldDIA、Skyline 和 Spectronaut 这 6 种不同的 DIA 数据分析软件。评估指标包括蛋白质鉴定数、重现性(coefficient of variation, CV)、蛋白质丰度和信号强度的线性度(coefficient of determination, R^2)以及蛋白质定量准确度(mean absolute percentage error, MAPE)等。评估结果表明，所有软件在窄窗口下能够定量更多的肽段和蛋白质。目前的以肽为中心的 DIA 鉴定软件在宽窗口下的鉴定和定量仍有提升空间。与 Navarro 等[41]的研究结果一致，现有 DIA 分析软件在低丰度蛋白质检测方面表现不佳。此外，他们的评估结果还证实了 DIA 在无标记定量方面具有高重现性，并揭示了蛋白质浓度与蛋白质定量结果之间的强相关性。DIA-NN 和 Spectronaut 的定量准确性较高，Skyline 容易受到噪声干扰，表现较差。在使用窄窗口下的质谱数据、预测谱图库以及 Spectronaut 软件时，蛋白质鉴定的灵敏度最高，其次是使用窄窗口采集的质谱数据、预测谱图库以及 DIA-NN 软件。

Fröhlich 等[40]构建了一套标准数据集，用于评估不同的 DIA 数据分析流程。其中数据分析使用了 4 种软件(DIA-NN、Skyline、OpenSWATH、Spectronaut)，谱图库有 3 种来源(GPF-DIA 建库、Prosit 预测谱图库和 DDA 实验建库)。评估指标包括蛋白质鉴定数、蛋白质丰度分布以及定量可重现性。评估结果表明，在谱图库选择方面，使用 DDA 谱图库鉴定到的蛋白质数最少，深度学习预测谱图库其次，而使用 GPF-DIA 谱图库鉴定到的蛋白质数最多；在分析软件选择方面，DIA-NN 鉴定到的蛋白质最多，结果最灵敏。

Lou等[43]构建了在Orbitrap仪器和timsTOF仪器上采集的DIA基准数据集，评估了4种常用的DIA软件(DIA-NN、Spectronaut、MaxDIA和Skyline)以及不同来源的谱图库(DDA实验建库、Prosit预测建库和DIA实验建库)共10条数据分析流程。评估指标包括了蛋白质组鉴定深度、谱图库质量对鉴定错误率和打分稳定性的影响、蛋白质组定量准确性和重复性。他们还利用一套合成磷酸化肽段的标注数据集评估了不同软件定位磷酸化修饰位点的准确性，并提出了适用于不同软件算法的修饰位点定位打分阈值。评估结果表明了DIA-NN和Spectronaut能在复杂样本中鉴定到更多的低丰度蛋白质，并且定量准确度和重复性更好。在磷酸肽的检测中，Spectronaut表现出更高的灵敏度和错误发现率，DIA-NN表现出更加严格的错误率控制，但牺牲了部分正确的磷酸化位点。

上述评估结果可供蛋白质组学领域的研究者在选择数据分析软件时参考。发布的数据集也可用于优化实验参数，评估新开发的软件工具。在包含数百到数千个样本的大规模DIA数据分析中，考虑到灵敏度和分析时间，研究者普遍认为以肽为中心的DIA数据分析方法是最佳选择[44]。在选择软件时，不仅要考虑软件的性能，还要考虑用户友好性、是否适用于计算机集群分析、是否利于下游数据分析、运行速度以及维护程度等因素。科研用软件DIA-NN、OpenSWATH以及商业软件Spectronaut都可并行运行，计算效率较高，适用于大样本量分析[45]。此外，DIA-NN提供了每一个样本的质控结果，有助于样本层面的质控[46]。在易用性方面，DIA-NN、Skyline、ScaffoldDIA都提供了图形用户界面，而OpenSWATH的命令行模式需要研究人员具备一定的编程能力。

3 改善以肽为中心DIA数据分析流程的研究进展

近年来，研究人员围绕以肽为中心的DIA数据分析流程，提出了诸多改进和优化方法。首先，在谱图库构建方面，谱图库的全面性、特异性和准确性直接影响DIA数据分析结果，因此提高谱图库的质量是非常重要的问题。Midha等[47]研发了谱图库质量控制工具DIALib-QC，计算了62种特征指标，用于系统评估谱图库质量。经过DIA-LibQC质控后的DDA谱图库可以提高30%的肽段鉴定数。Isaksson等[48]发表了MSLibrarian，该工具基于Prosit、DIA-Umpire以及DeepLC来优化谱图库：首先使用DIA-Umpire分析DIA数据，然后通过实际鉴定谱图和对应理论谱图的相似性优化Prosit预测谱图的参数，并校正DeepLC预测的保留时间，实现对谱图库的优化。作者使用优化的谱图库提高了8.3%的肽段鉴定量和5.7%的蛋白质鉴定量。Ge等[49]提出了优化谱图库的计算策略subLib，该方法首先利用FDR 5%的条件下产生的第一次鉴定结果生成子谱图库，然后将子谱图库应用于二次鉴定。作者使用该工具优化后的谱图库，多鉴定了39.2%的肽段和19%的蛋白质。Yang等[23]使用深度学习模型预测肽段可检测性(肽段离子在质谱中被检测到的概率)，过滤掉谱图库中可检测性较低的肽段，可一定程度上提高蛋白质鉴定种类约2.3%，并且降低了鉴定结果中的假阳性率。此外，本文作者使用肽段酶切概率(digestibility)过滤谱图库也可以提高鉴定的灵敏度，使用了2021年Yang等[50]发表在*Analytical Chemistry*上的DeepDigest算法预测谱图库中肽段的酶切概率，并通过设定酶切概率的最小阈值来过滤谱图库(包含20 368个蛋白质)中的肽段。过滤之后

的谱图库按照 1:1 加入陷阱[51]肽段，用于评估鉴定结果的假阳性。采用 LFQbench[41]的 8 个子数据集(每个子数据集是由不同的实验仪器在不同的采集窗口大小下产生的，包含了 6 个质谱原始文件)评估酶切概率的过滤效果，采用 DIA-NN 进行 DIA 数据分析。结果表明，去除谱图库中肽段酶切概率低的肽段可以有效提高蛋白质鉴定数约 2.8%，并减少结果中的假阳性匹配(图 3)。

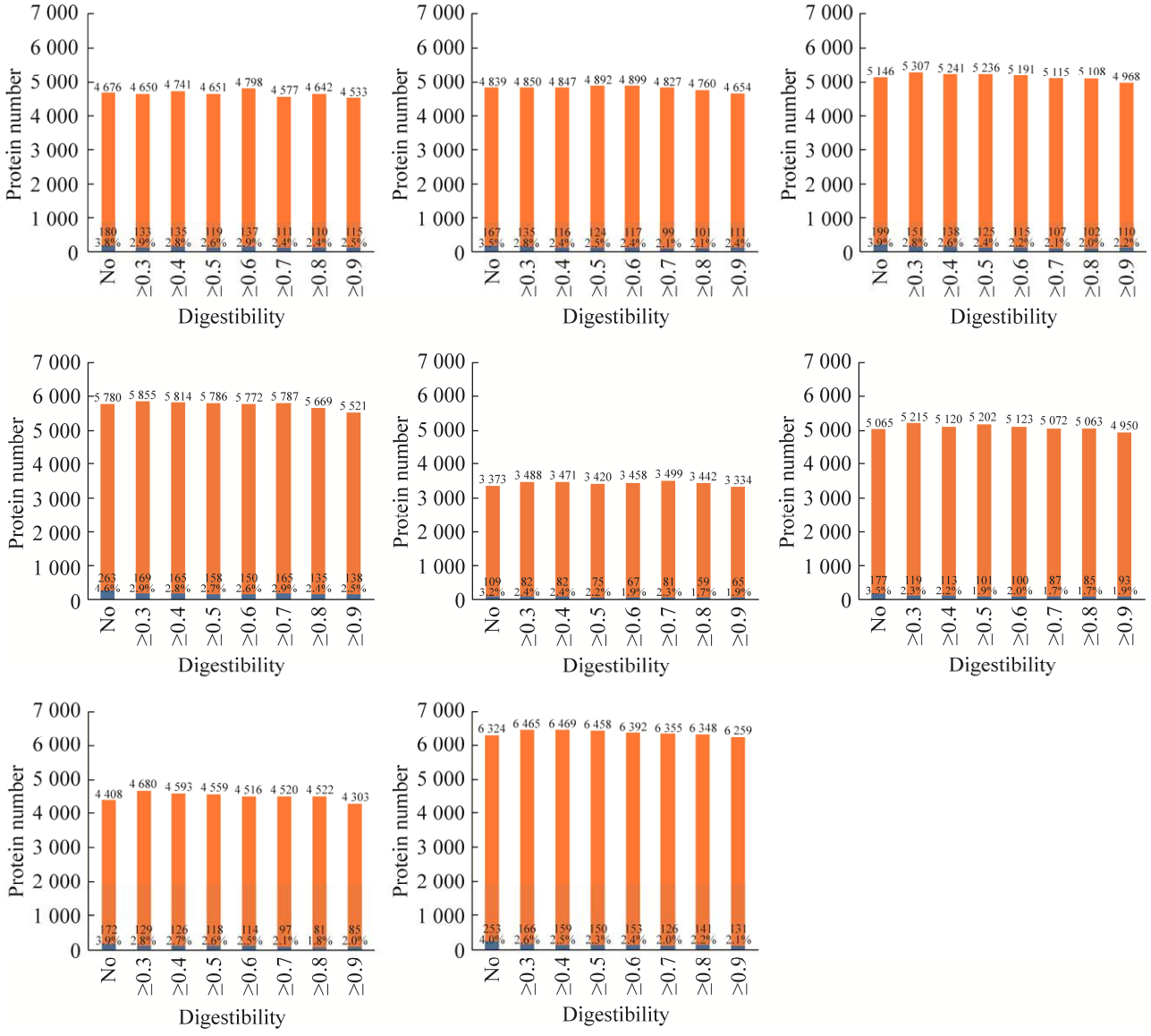

图 3　LFQbench 8 个子数据集的蛋白质鉴定数　橙色代表人源蛋白质的鉴定数目，蓝色代表陷阱蛋白质的鉴定数目. 百分数代表陷阱蛋白占总鉴定数的比例. 横轴代表不同的酶切概率阈值，纵轴代表蛋白质鉴定数目

Figure 3　The number of identified proteins from eight datasets of LFQbench. Orange and blue are the identification number of human proteins and entrapment proteins, respectively. The percentages of entrapment proteins using different digestibility cutoffs are lised in each plot. *X*-axis is the digestibility cutoff and *Y*-axis is the number of the indentified proteins.

其次在特征打分方面，由于深度学习拥有强大的特征学习功能，越来越多的研究开始采用神经网络学习模型，利用数据驱动方式计算特征打分。Song 等发表了 Alpha-XIC[52] 和 Alpha-Tri[53]，利用神经网络打分离子共流出和强度相似性的特征，并与 DIA-NN 的其他特征结合，提高了 DIA-NN 的鉴定灵敏度。Gao 等[35] 提出了 DreamDIA，设计了新型谱图数据结构代表性谱图特征(representative spectral matrix, RSM)，包括谱图库离子色谱峰、理论碎片离子色谱峰以及同位素峰等 6 种不同类型的色谱峰信息。通过使用长短时记忆(long short-term memory, LSTM)网络提取低维深度表示特征，并与其他特征相结合构建非线性判别模型来计算 FDR，从而实现了肽段的精准鉴定。

最后，在统计质控方面，由于 DIA 数据高度复杂，DIA 数据的质控要求也更高。当前使用的目标诱饵竞争的质控策略可能会导致鉴定结果中的假阴性和假阳性过高。一方面，如果谱图库中的阴性肽段数量过多，目标肽段和诱饵肽段难以区分，会导致更严格的统计质控，降低了蛋白质鉴定的灵敏度，增加了假阴性的数量。另一方面，现有以肽为中心，靶向提取质谱信号的数据分析方法中，检测到诱饵肽的概率与检测到目标肽中假阳性匹配的概率并不完全相同[54]。比如，生成的诱饵肽和目标肽中漏切位点数不一致，在鉴定软件中诱饵肽段的得分会显著降低。诱饵“欺骗”鉴定软件的程度不够，导致鉴定结果中假阳性升高[55]。Rosenberger 等[46]提出使用包含大量阴性肽段的谱图库时，不仅要在肽段水平上进行错误率控制，也要在肽谱匹配、蛋白质的水平上进行错误率控制。在进行多个样本的全局质控时，由于不同样本数据组间存在异质性，可能导致某些样本的 FDR 估计偏低，而其他样本则偏高。Freestone 等[56]提出了 Group-walk 策略，旨在分析目标-诱饵竞争策略的同时利用实验的不同分组进行质控，解决组间异质性的问题。

以上的改进工作表明，DIA 数据分析流程的每个步骤仍有待进一步完善，为算法研究者提供了理论基础和改进方向。

4 总结与展望

尽管以肽为中心的 DIA 数据分析方法已经在蛋白质组学研究中成功应用，但仍存在诸多问题亟待解决：(1) 构建谱图库对以肽为中心的数据分析方法至关重要。DDA 实验建库需要较长的时间成本，并且其包含的蛋白质种类有限。使用蛋白质序列数据库预测建库时，虽然节约了时间，提高了谱图的完整性，但由于谱图库中阴性肽段过多，降低了鉴定灵敏度。谱图库特异性和完整性的平衡问题仍然需要进一步研究。本文提出了一种基于酶切概率过滤谱图库的新方法，可以提高谱图库的特异性，从而在一定程度上提高了蛋白质鉴定的灵敏度。(2) 作为以肽为中心数据分析算法的关键步骤之一，根据谱图库靶向提取离子信号直接影响着 DIA 数据分析的鉴定灵敏度和定量准确度。一方面，色谱峰提取算法需要足够高的准确性和灵敏性去检测低丰度肽段信号和区分噪声和信号。另一方面，进行大队列蛋白质组数据分析时，由于质谱仪的稳定性有限，肽段的 RT 和 *m*/*z* 可能会产生随机偏移，这会导致样本间具有一定比例的缺失值，从而影响肽段、蛋白质的鉴定和定量。因此，需要准确有效的 RT 和 *m*/*z* 校正算法，以及不同样本之间 RT 对齐算法，来提高鉴定的准确性和灵敏性。(3) 对于色谱峰群的特征打分，仍然缺乏全面、系统的打分方法。(4) 在谱图库规模和数据复杂度增

大的情况下，在 DDA 数据上发展而来的目标诱饵竞争策略并不适用于 DIA 数据的质控分析，因此有必要深入研究适用于 DIA 数据的质控方法。

DIA 技术将在蛋白质组学领域，特别是在基于大队列的临床蛋白质组学研究中发挥不可替代的重要作用。学术界亟需开发精准、高效的 DIA 数据分析算法及软件。本文希望为从事 DIA 数据分析、算法和软件研发的科研人员提供方法学参考。

REFERENCES

[1] WILKINS MR, APPEL RD. Ten Year of the Proteome[M]. Berlin: Springer Berlin Heidelberg, 2007: 1-13.

[2] WILHELM M, SCHLEGL J, HAHNE H, GHOLAMI AM, LIEBERENZ M, SAVITSKI MM, ZIEGLER E, BUTZMANN L, GESSULAT S, MARX H, MATHIESON T, LEMEER S, SCHNATBAUM K, REIMER U, WENSCHUH H, MOLLENHAUER M, SLOTTA-HUSPENINA J, BOESE JH, BANTSCHEFF M, GERSTMAIR A, et al. Mass-spectrometry-based draft of the human proteome[J]. Nature, 2014, 509(7502): 582-587.

[3] ADHIKARI S, NICE EC, DEUTSCH EW, LANE L, OMENN GS, PENNINGTON SR, PAIK YK, OVERALL CM, CORRALES FJ, CRISTEA IM, van EYK JE, UHLÉN M, LINDSKOG C, CHAN DW, BAIROCH A, WADDINGTON JC, JUSTICE JL, LABAER J, RODRIGUEZ H, HE FC, et al. A high-stringency blueprint of the human proteome[J]. Nature Communications, 2020, 11(1): 1-16.

[4] 孙瑞祥, 付岩, 李德泉, 张京芬, 王晓彪, 盛泉虎, 曾嵘, 陈益强, 贺思敏, 高文. 基于质谱技术的计算蛋白质组学研究[J]. 中国科学 E 辑: 信息科学, 2006, 36(2): 222-234.
SUN RX, FU Y, LI DQ, ZHANG JF, WANG XB, SHENG QH, ZENG R, CHEN YQ, HE SM, GAO W. Study on computational protein genomics based on mass spectrometry[J]. Science in China (Series E: Information Sciences), 2006, 36(2): 222-234 (in Chinese).

[5] VENABLEJD, DONG MQ, WOHLSCHLEGEL J, DILLIN A, YATES JR III. Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra[J]. Nature Methods, 2004, 1(1): 39-45.

[6] 王震. 基于质谱 DIA 数据处理算法软件评估与肽段定量信号提取算法研究[D]. 长沙: 国防科学技术大学硕士学位论文, 2016.
WANG Z. Algorithms evaluation based on data-independence-acquisition data processing and research on quantitative signal extraction algorithm for peptide[D]. Changsha: Master's Thesis of National University of Defense Technology, 2016 (in Chinese).

[7] PURVINE S, EPPEL JT, YI EC, GOODLETT DR. Shotgun collision-induced dissociation of peptides using a time of flight mass analyzer[J]. Proteomics, 2003, 3(6): 847-850.

[8] GEIGER T, COX J, MANN M. Proteomics on an orbitrap benchtop mass spectrometer using all-ion fragmentation[J]. Molecular & Cellular Proteomics, 2010, 9(10): 2252-2261.

[9] ROBERT SP, KELLY AJ, PAUL R, BRIAN WS, IAN DW, JOSE MC, JEREMY KN. UPLC/MS^{E}; a new approach for generating molecular fragment information for biomarker structure elucidation[J]. Rapid Communications in Mass Spectrometry, 2006, 20(14): 2234-2234.

[10] GILLET LC, NAVARRO P, TATE S, RÖST H, SELEVSEK N, REITER L, BONNER R, AEBERSOLD R. Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis[J]. Molecular & Cellular Proteomics, 2012, 11(6): 1-17.

[11] EGERTSON JD, KUEHN A, MERRIHEW GE, BATEMAN NW, MACLEAN BX, TING YS, CANTERBURY JD, MARSH DM, KELLMANN M, ZABROUSKOV V, WU CC, MACCOSS MJ. Multiplexed MS/MS for improved data-independent acquisition[J]. Nature Methods, 2013, 10(8): 744-746.

[12] MOSELEY MA, HUGHES CJ, JUVVADI PR, SODERBLOM EJ, LENNON S, PERKINS SR, THOMPSON JW, STEINBACH WJ, GEROMANOS SJ, WILDGOOSE J, LANGRIDGE JI, RICHARDSON K, VISSERS JPC. Scanning quadrupole data-independent acquisition, part A: qualitative and quantitative characterization[J]. Journal of Proteome Research, 2018, 17(2): 770-779.

[13] BRUDERER R, BERNHARDT OM, GANDHI T,

MILADINOVIĆ SM, CHENG LY, MESSNER S, EHRENBERGER T, ZANOTELLI V, BUTSCHEID Y, ESCHER C, VITEK O, RINNER O, REITER L. Extending the limits of quantitative proteome profiling with data-independent acquisition and application to acetaminophen-treated three-dimensional liver microtissues[J]. Molecular & Cellular Proteomics, 2015, 14(5): 1400-1410.

[14] ZHANG Y, BILBAO A, BRUDERER T, LUBAN J, STRAMBIO-de-CASTILLIA C, LISACEK F, HOPFGARTNER G, VARESIO E. The use of variable Q1 isolation windows improves selectivity in LC-SWATH-MS acquisition[J]. Journal of Proteome Research, 2015, 14(10): 4359-4371.

[15] MEIER F, GEYER PE, VIRREIRA WINTER S, COX J, MANN M. BoxCar acquisition method enables single-shot proteomics at a depth of 10 000 proteins in 100 minutes[J]. Nature Methods, 2018, 15(6): 440-448.

[16] LUDWIG C, GILLET L, ROSENBERGER G, AMON S, COLLINS BC, AEBERSOLD R. Data-independent acquisition-based SWATH-MS for quantitative proteomics: a tutorial[J]. Molecular Systems Biology, 2018, 14(8): 1-23.

[17] ROSENBERGER G, KOH CC, GUO TN, RÖST HL, KOUVONEN P, COLLINS BC, HEUSEL M, LIU YS, CARON E, VICHALKOVSKI A, FAINI M, SCHUBERT OT, FARIDI P, ALEXANDER EBHARDT H, MATONDO M, LAM H, BADER SL, CAMPBELL DS, DEUTSCH EW, MORITZ RL, et al. A repository of assays to quantify 10 000 human proteins by SWATH-MS[J]. Scientific Data, 2014, 1: 1-15.

[18] ZHU TS, ZHU Y, XUAN Y, GAO HH, CAI X, PIERSMA SR, PHAM TV, SCHELFHORST T, HAAS RRGD, BIJNSDORP IV, SUN R, YUE L, RUAN G, ZHANG QS, HU M, ZHOU Y, van HOUDT WJ, Le LARGE TYS, CLOOS J, WOJTUSZKIEWICZ A, et al. DPHL: a DIA pan-human protein mass spectrometry library for robust biomarker discovery[J]. Genomics, Proteomics & Bioinformatics, 2020, 18(2): 104-119.

[19] ZHOU XX, ZENG WF, CHI H, LUO CJ, LIU C, ZHAN JF, HE SM, ZHANG ZF. pDeep: predicting MS/MS spectra of peptides with deep learning[J]. Analytical Chemistry, 2017, 89(23): 12690-12697.

[20] ZOLG DANIEL P, MATHIAS W, KARSTEN S, JOHANNES Z, TOBIAS K, BERNARD D, BAILEY DEREK J, SIEGFRIED G, HANS-CHRISTIAN E, MAXIMILIAN W, PENG Y, JUDITH S, KARL K, TOBIAS S, ULRIKE K, DEUTSCH ERIC W, RUEDI A, MORITZ ROBERT L, HOLGER W, THOMAS M, et al. Building ProteomeTools based on a complete synthetic human proteome[J]. Nature Methods, 2017, 14(3): 259-262.

[21] MA CW, REN Y, YANG JR, REN Z, YANG HM, LIU SQ. Improved peptide retention time prediction in liquid chromatography through deep learning[J]. Analytical Chemistry, 2018, 90(18): 10881-10888.

[22] GESSULAT S, SCHMIDT T, ZOLG DP, SAMARAS P, SCHNATBAUM K, ZERWECK J, KNAUTE T, RECHENBERGER J, DELANGHE B, HUHMER A, REIMER U, EHRLICH HC, AICHE S, KUSTER B, WILHELM M. Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning[J]. Nature Methods, 2019, 16(6): 509-518.

[23] YANG Y, LIU XH, SHEN CP, LIN Y, YANG PY, QIAO L. *In silico* spectral libraries by deep learning facilitate data-independent acquisition proteomics[J]. Nature Communications, 2020, 11(1): 146.

[24] BOUWMEESTER R, GABRIELS R, HULSTAERT N, MARTENS L, DEGROEVE S. DeepLC can predict retention times for peptides that carry as-yet unseen modifications[J]. Nature Methods, 2021, 18(11): 1363-1369.

[25] WU JX, SONG XM, PASCOVICI D, ZAW T, CARE N, KRISP C, MOLLOY MP. SWATH mass spectrometry performance using extended peptide MS/MS assay libraries[J]. Molecular & Cellular Proteomics, 2016, 15(7): 2501-2514.

[26] PINO LK, JUST SC, MACCOSS MJ, SEARLE BC. Acquiring and analyzing data independent acquisition proteomics experiments without spectrum libraries[J]. Molecular & Cellular Proteomics, 2020, 19(7): 1088-1103.

[27] JÜRGEN C, MATTHIAS M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification[J]. Nature Biotechnology, 2008, 26(12): 1367-1372.

[28] RÖST HL, ROSENBERGER G, NAVARRO P, GILLET L, MILADINOVIĆ SM, SCHUBERT OT, WOLSKI W, COLLINS BC, MALMSTRÖM J, MALMSTRÖM L, AEBERSOLD R. OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data[J]. Nature Biotechnology, 2014, 32(3): 219-223.

[29] DEMICHEV V, MESSNER CB, VERNARDIS SI,

LILLEY KS, RALSER M. DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput[J]. Nature Methods, 2020, 17(1): 41-44.

[30] ESCHER C, REITER L, MACLEAN B, OSSOLA R, HERZOG F, CHILTON J, MACCOSS MJ, RINNER O. Using iRT, a normalized retention time for more targeted measurement of peptides[J]. Proteomics, 2012, 12(8): 1111-1121.

[31] SINITCYN P, HAMZEIY H, SALINAS SOTO F, ITZHAK D, MCCARTHY F, WICHMANN C, STEGER M, OHMAYER U, DISTLER U, KASPAR-SCHOENEFELD S, PRIANICHNIKOV N, YıLMAZ Ş, RUDOLPH JD, TENZER S, PEREZ-RIVEROL Y, NAGARAJ N, HUMPHREY SJ, COX J. MaxDIA enables library-based and library-free data-independent acquisition proteomics[J]. Nature Biotechnology, 2021, 39(12): 1563-1573.

[32] ELIAS JE, GYGI SP. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry[J]. Nature Methods, 2007, 4(3): 207-214.

[33] REITER L, RINNER O, PICOTTI P, HÜTTENHAIN R, BECK M, BRUSNIAK MY, HENGARTNER MO, AEBERSOLD R. mProphet: automated data processing and statistical validation for large-scale SRM experiments[J]. Nature Methods, 2011, 8(5): 430-435.

[34] TELEMAN J, RÖST HL, ROSENBERGER G, SCHMITT U, MALMSTRÖM L, MALMSTRÖM J, LEVANDER F. DIANA—algorithmic improvements for analysis of data-independent acquisition MS data[J]. Bioinformatics, 2015, 31(4): 555-562.

[35] GAO MX, YANG WX, LI CX, CHANG YQ, LIU YC, HE QZ, ZHONG CQ, SHUAI JW, YU RS, HAN JH. Deep representation features from DreamDIA^{XMBD} improve the analysis of data-independent acquisition proteomics[J]. Communications Biology, 2021, 4(1): 1-10.

[36] SEARLE BC, PINO LK, EGERTSON JD, TING YS, LAWRENCE RT, MACLEAN BX, VILLÉN J, MACCOSS MJ. Chromatogram libraries improve peptide detection and quantification by data independent acquisition mass spectrometry[J]. Nature Communications, 2018, 9(1): 5128.

[37] BERNHARDT OM, SELEVSEK N, GILLET LC, RINNER O, PICOTTI P, AEBERSOLD R, REITER L. Spectronaut: a Fast and Efficient Algorithm for MRM-Like Processing of Data Independent Acquisition (SWATH-MS) Data[C]. 60th American Society for Mass Spectometry Conference, 2014.

[38] MACLEAN B, TOMAZELA DM, SHULMAN N, CHAMBERS M, FINNEY GL, FREWEN B, KERN R, TABB DL, LIEBLER DC, MACCOSS MJ. Skyline: an open source document editor for creating and analyzing targeted proteomics experiments[J]. Bioinformatics, 2010, 26(7): 966-968.

[39] GOTTI C, ROUX-DALVAI F, JOLY-BEAUPARLANT C, MANGNIER L, LECLERCQ M, DROIT A. Extensive and accurate benchmarking of DIA acquisition methods and software tools using a complex proteomic standard[J]. Journal of Proteome Research, 2021, 20(10): 4801-4814.

[40] FRÖHLICH K, BROMBACHER E, FAHRNER M, VOGELE D, KOOK L, PINTER N, BRONSERT P, TIMME-BRONSERT S, SCHMIDT A, BÄRENFALLER K, KREUTZ C, SCHILLING O. Benchmarking of analysis strategies for data-independent acquisition proteomics using a large-scale dataset comprising inter-patient heterogeneity[J]. Nature Communications, 2022, 13(1): 2622.

[41] NAVARRO P, KUHAREV J, GILLET LC, BERNHARDT OM, MACLEAN B, RÖST HL, TATE SA, TSOU CC, REITER L, DISTLER U, ROSENBERGER G, PEREZ-RIVEROL Y, NESVIZHSKII AI, AEBERSOLD R, TENZER S. A multicenter study benchmarks software tools for label-free proteome quantification[J]. Nature Biotechnology, 2016, 34(11): 1130-1136.

[42] TSOU CC, AVTONOMOV D, LARSEN B, TUCHOLSKA M, CHOI H, GINGRAS AC, NESVIZHSKII AI. DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics[J]. Nature Methods, 2015, 12(3): 258-264.

[43] LOU RH, CAO Y, LI SS, LANG XY, LI YX, ZHANG YY, SHUI WQ. Benchmarking commonly used software suites and analysis workflows for DIA proteomics and phosphoproteomics[J]. Nature Communications, 2023, 14(1): 94.

[44] COLLINS BEN C, HUNTER CHRISTIE L, LIU YS, BIRGIT S, GEORGE R, BADER SAMUEL L, CHAN DANIEL W, GIBSON BRADFORD W, ANNE-CLAUDE G, HELD JASON M, MIO HK, HOU GX, CHRISTOPH K, BRETT L, LIN L, LIU SQ, MOLLOY MARK P, MORITZ ROBERT L, SUMIO O, RALPH S, et al. Multi-laboratory assessment of

reproducibility, qualitative and quantitative performance of SWATH-mass spectrometry[J]. Nature Communications, 2017, 8(1): 291.

[45] ZHANG FF, GE WG, RUAN G, CAI X, GUO TN. Data-independent acquisition mass spectrometry-based proteomics and software tools: a glimpse in 2020[J]. Proteomics, 2020, 20(17/18): 1-12.

[46] ROSENBERGER G, BLUDAU I, SCHMITT U, HEUSEL M, HUNTER CL, LIU YS, MACCOSS MJ, MACLEAN BX, NESVIZHSKII AI, PEDRIOLI PGA, REITER L, RÖST HL, TATE S, TING YS, COLLINS BC, AEBERSOLD R. Statistical control of peptide and protein error rates in large-scale targeted data-independent acquisition analyses[J]. Nature Methods, 2017, 14(9): 921-927.

[47] MIDHA MK, CAMPBELL DS, KAPIL C, KUSEBAUCH U, HOOPMANN MR, BADER SL, MORITZ RL. DIALib-QC an assessment tool for spectral libraries in data-independent acquisition proteomics[J]. Nature Communications, 2020, 11(1): 1-8.

[48] ISAKSSON M, KARLSSON C, LAURELL T, KIRKEBY A, HEUSEL M. MSLibrarian: optimized predicted spectral libraries for data-independent acquisition proteomics[J]. Journal of Proteome Research, 2022, 21(2): 535-546.

[49] GE WG, LIANG X, ZHANG FF, HU YF, XU LA, XIANG N, SUN R, LIU W, XUE ZZ, YI X, SUN YT, WANG B, ZHU J, LU C, ZHAN XL, CHEN LR, WU Y, ZHENG ZG, GONG WG, WU QJ, et al. Computational optimization of spectral library size improves DIA-MS proteome coverage and applications to 15 tumors[J]. Journal of Proteome Research, 2021, 20(12): 5392-5401.

[50] YANG JH, GAO ZQ, REN XH, SHENG J, XU P, CHANG C, FU Y. DeepDigest: prediction of protein proteolytic digestion with deep learning[J]. Analytical Chemistry, 2021, 93(15): 6094-6103.

[51] FENG XD, LI LW, ZHANG JH, ZHU YP, CHANG C, SHU KX, MA J. Using the entrapment sequence method as a standard to evaluate key steps of proteomics data analysis process[J]. BMC Genomics, 2017, 18(2): 1-9.

[52] SONG J, YU CB. Alpha-XIC: a deep neural network for scoring the coelution of peak groups improves peptide identification by data-independent acquisition mass spectrometry[J]. Bioinformatics, 2021, 38(1): 38-43.

[53] SONG J, YU CB. Alpha-Tri: a deep neural network for scoring the similarity between predicted and measured spectra improves peptide identification of DIA data[J]. Bioinformatics, 2022, 38(6): 1525-1531

[54] COUTÉ Y, BRULEY C, BURGER T. Beyond target-decoy competition: stable validation of peptide and protein identifications in mass spectrometry-based discovery proteomics[J]. Analytical Chemistry, 2020, 92(22): 14898-14906.

[55] DANILOVA Y, VORONKOVA A, SULIMOV P, KERTÉSZ-FARKAS A. Bias in false discovery rate estimation in mass-spectrometry-based peptide identification[J]. Journal of Proteome Research, 2019, 18(5): 2354-2358.

[56] FREESTONE J, SHORT T, STAFFORD NOBLE W, KEICH U. Group-walk: a rigorous approach to group-wise false discovery rate analysis by target-decoy competition[J]. Bioinformatics, 2022, 38(supplement_2): ii82-ii88.

(本文责编　陈宏宇)