PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2022,49(12):2364~2386
www.pibb.ac.cn

基于数据非依赖采集的蛋白质组质谱数据解析方法研究进展*

侯鑫行[1)] 周丕宇[1)] 宫鹏云[2)] 付嘉乐[3)] 刘 超[2)**] 王海鹏[1)**]

([1)] 山东理工大学计算机科学与技术学院，淄博 255000；
[2)] 北京航空航天大学医学科学与工程学院&生物与医学工程学院，北京 100191；
[3)] 清华大学生命科学学院，北京 100084）

摘要 数据非依赖采集（DIA）是蛋白质组学领域近年来快速发展的质谱采集技术，其通过无偏碎裂隔离窗口内的所有母离子采集二级谱图，理论上可实现蛋白质样品的深度覆盖，同时具有高通量、高重现性和高灵敏度的优点。现有的DIA数据采集方法可以分为全窗口碎裂方法、隔离窗口序列碎裂方法和四维DIA数据采集方法（4D-DIA）3大类。针对DIA数据的不同特点，主要数据解析方法包括谱库搜索方法、蛋白质序列库直接搜索方法、伪二级谱图鉴定方法和从头测序方法4大类。解析得到的肽段鉴定结果需要进行可信度评估，包括使用机器学习方法的重排序和对报告结果集合的假发现率估计两个步骤，实现对数据解析结果的质控。本文对DIA数据的采集方法、数据解析方法及软件和鉴定结果可信度评估方法进行了整理和综述，并展望了未来的发展方向。

关键词 蛋白质组学，数据非依赖采集，质谱，靶向数据提取
中图分类号 Q51，TP39 **DOI**：10.16476/j.pibb.2021.0345

基于液相色谱-串联质谱技术的鸟枪法蛋白质组学已成为对生物体内蛋白质进行全面分析的一个主流方法[1-2]，在疾病发病机制研究、生物标志物筛选和药靶发现等领域有着广泛的应用[3-7]。目前，鸟枪法蛋白质组学已实现对大规模生物样品上万个蛋白质的高通量分析[7-12]，但依然还未实现蛋白质全覆盖的目标[13]。传统鸟枪法蛋白质组学通常采用数据依赖采集（data dependent acquisition，DDA）方法采集质谱数据，而近年提出的数据非依赖采集（data independent acquisition，DIA）方法由于其高通量、高重现性和高灵敏度的优点，逐渐得到更广泛的应用。

DDA方法在每次循环采集过程中依次选择一级谱图中强度最高的多个母离子，对母离子质荷比选择较窄的隔离窗口（如±1.2 u）进行碎裂并生成二级谱图（图1a）。DDA方法依赖母离子强度的二级谱图获取方式导致其难以获取低丰度肽段的二级谱图，并且由于色谱条件和动态排除机制造成了母离子选择的高随机性，使得相同肽段在两次实验被重复采集二级谱图的概率较低。

随着质谱技术的不断发展，为了实现对蛋白质样品的高通量、高灵敏和高重现性分析，研究人员提出了DIA方法[14]。DIA方法不依赖于母离子强度，而是根据母离子质荷比范围无偏地设置隔离窗口，并将窗口内所有母离子共碎裂，得到包含多个母离子共碎裂信息的混合二级谱图（图1b）。相较于DDA方法，DIA方法可以采集到包括低丰度肽段在内隔离窗口中所有母离子的碎片离子信息，理论上可实现对肽段的全面采集，获取样品内全部蛋白质的完整图谱。此外，DIA方法采集了碎片离子在色谱时间上的连续信息，可以重构碎片离子的色谱曲线，该色谱曲线较DDA方法中重构母离子色

* 山东省高等学校优秀青年创新团队支持计划（2019KJN048）和国家自然科学基金（31500669）资助项目。
** 通讯联系人。
王海鹏 Tel：0533-2783479，E-mail：hpwang@sdut.edu.cn
刘超 Tel：010-82316427，E-mail：liuchaobuaa@buaa.edu.cn
收稿日期：2021-11-16，接受日期：2022-03-21

谱曲线有更高的信噪比和更低的检测限，能够实现更精确的定量。

DIA数据虽然具有对样品所有肽段进行鉴定和定量的潜能，但是高度复杂的混合二级谱图对肽段和蛋白质的准确鉴定提出了挑战。由于DIA产生的二级谱图包含着隔离窗口所有母离子的碎裂信息，母离子和碎片离子之间的对应关系被打破，难以直接使用传统DDA搜索引擎实现肽段鉴定。此外，隔离窗口内多个母离子之间存在相同质荷比的碎片离子，造成碎片离子干扰为二级谱图解析造成困难。自DIA方法提出以来，正确解析DIA数据的二级谱图成为了分析DIA数据的关键难点。

近年来，质谱采集技术和不同鉴定策略的发展为解析DIA数据的二级谱图提供了有效途径。随着质谱仪器在质量精度、速度和分辨率上的提高，多种旨在降低分析复杂度的DIA采集策略被提出。目前的DIA数据采集方法主要可以分为全窗口碎裂方法（Shotgun-CID[15]、MS^E[16]、AIF[17]）、隔离窗口序列碎裂方法（原始DIA[14]、PAcIFIC[18]、SWATH[19]、WiSIM-DIA[20]、BoxCar[21]、MSX[22]、可变母离子隔离窗口DIA[23]、RTWinDIA[24]）和增加数据维度的4D-DIA方法（DIA-PASEF[25]、ScanningSWATH[26]）。

针对DIA数据的特点，基于不同策略的数据分析方法被提出。DIA数据分析主要包括数据解析获取肽段鉴定结果（即实现肽段和谱图匹配，简称肽谱匹配，peptide-spectrum matches）和对鉴定结果进行可信度评估两个部分。目前，DIA数据解析方法主要包括谱库搜索方法（mProphet[27]、OpenSWATH[28]、Spectronaut[29]、DIANA[30]、SWATHProphet[31]、EncyclopeDIA[32]、DIA-NN[33]、DDIA[34]、MaxDIA[35]、DreamDIA[36]）、蛋白质序列库直接搜索方法（DIAmeter[37]、FT-ARM[38]、PECAN[39]）和伪二级谱图鉴定方法（DeMux[40]、DIA-Umpire[41]、Group-DIA[42]、Specter[43]、CorrDec[44]）以及从头测序方法（DeepNovo-DIA[45]）。在得到肽段鉴定结果后，使用重排序算法将鉴定结果按可信度从高到低进行排序，而后利用目标-诱饵库方法（target decoy approach，TDA）[46]估计假发现率（false discovery rate，FDR）[47]，实现对鉴定结果的可信度评估。

本文对近年来发展出的DIA数据分析策略进行综述。首先介绍主要的DIA数据采集方法，接着介绍主要的DIA数据解析方法，然后介绍DIA数据中的鉴定结果重排序算法和假发现率估计方法，最后对现有蛋白质组学中的DIA分析策略进行总结并对未来发展进行展望。

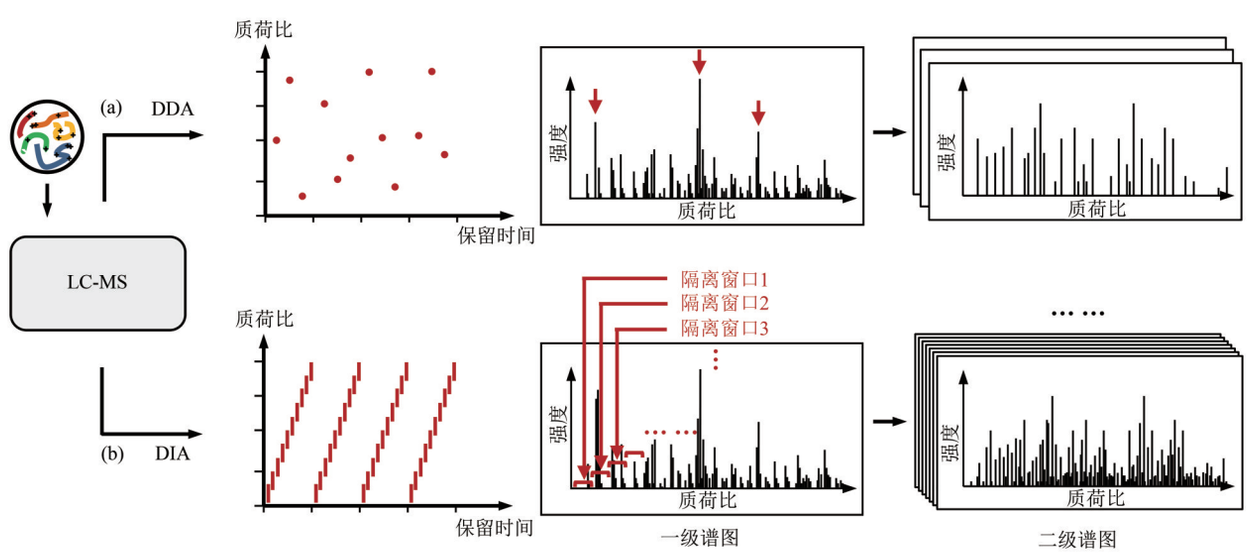


Fig. 1　Comparison of DDA and DIA methods

图1　DDA和DIA方法对比

1 DIA数据采集

DIA数据采集方法对一级谱图的母离子质荷比范围进行划分，得到隔离窗口并对隔离窗口内所有母离子共碎裂采集二级谱图。目前已经发展了多种DIA采集方法（表1），根据隔离窗口划分数目、大小的不同及是否对肽段采集额外的维度，主要分为3类，包括全窗口碎裂方法、隔离窗口序列碎裂方法和增加数据维度的四维DIA数据采集方法（4D-DIA）（图2）。

Table 1 Commonly used data-independent acquisition methods

表1 常用的数据非依赖采集方法

方法名称	发表年份	母离子范围（*m*/*z*）	隔离窗口大小（*m*/*z*）	二级谱图数目	循环时间/s	仪器	特点
Shotgun-CID	2003	Full	Full	1	2	Q-TOF	提出用母离子-碎片离子色谱曲线一致性辅助鉴定
原始DIA	2004	400~1 400	10	100	25~30	Thermo Electron LTQ	使用固定序列隔离窗口遍历母离子范围，提高数据重现性
MS^E	2006	Full	Full	1	2	Q-TOF	低能量和高能量交互碎裂循环产生一二级谱图
PAcIFIC	2009	400~1 400	2.5	10（每次实验）	~3	LTQ-Obritrap	利用GPF多次小隔离窗口实验覆盖母离子范围
AIF	2010	300~1 600	1 300	1	2	Orbitrap Exactive	使用步进式碎裂能量循环产生一二级谱图
SWATH-MS	2012	400~1 200	25	32	3.3	TripleTOF	使用重叠隔离窗后步进式遍历母离子范围
MSX	2013	500~900	20（4×5）	20	~3.5	Q-Extactive	多路窗口混合方法，随机选取5个小隔离窗口生成混合谱图
WiSim-DIA	2014	400~1 000	12	17×3	3.6	Orbitrap Fusion	将母离子范围划分为3个区间，每个区间独立采集一级和二级谱图
Variable window-DIA	2015	400~1 200	动态变化	32	3.3	TripleTOF	根据母离子分布情况（PIP）和总的离子流强度（TIC）来动态划分母离子隔离窗口范围
BoxCarDIA	2018	350~1 400	随质荷比变化	20	/	Orbitrap Fusion	对母离子质荷比范围划分成多个相隔的独立分配最大离子注入时间小窗口采集母离子信号
Overlapping Windows DIA	2019	500~900	20	20	~2.5	Q-Extactive	采用10 u的大比例重叠窗口，交替进行母离子范围的循环采集
RTwinDIA	2019	随时间变化	5	40	/	Orbitrap Fusion	根据保留时间改变采集的母离子范围，使用小窗口采集
DIA-PASEF	2019	400~1 200	25	16×2	3.3	TIMS TOF Pro	离子淌度分离与四极杆质量同步选择，添加离子淌度维度信息
HRMS1-DIA	2020	100~1 210	15	54	5.2	Q-Extactive HF	每次循环采集3个完整一级谱图，提高母离子信号数目重构色谱曲线，使用一级谱图定量分析
Pulse-DIA	2021	400~1 200	随质荷比变化	24（每次实验）	/	Q-Extactive HF-X	利用GPF将隔离窗口均匀划分为多份，分配到每个实验独立采集
BoxCarmax	2021	357~1 197	10（2.5×4）	30	/	Orbitrap Fusion	结合了BoxCar的母离子高灵敏度和MSX的母离子高选择性
Scanning SWATH	2021	400~900	10	/	0.52	Triple TOF	利用四级杆的连续扫描功能连续采集二级谱图，并累加碎片离子强度到bin中，增加Q1离子强度信息

/：文献中没有介绍或者不是固定值。

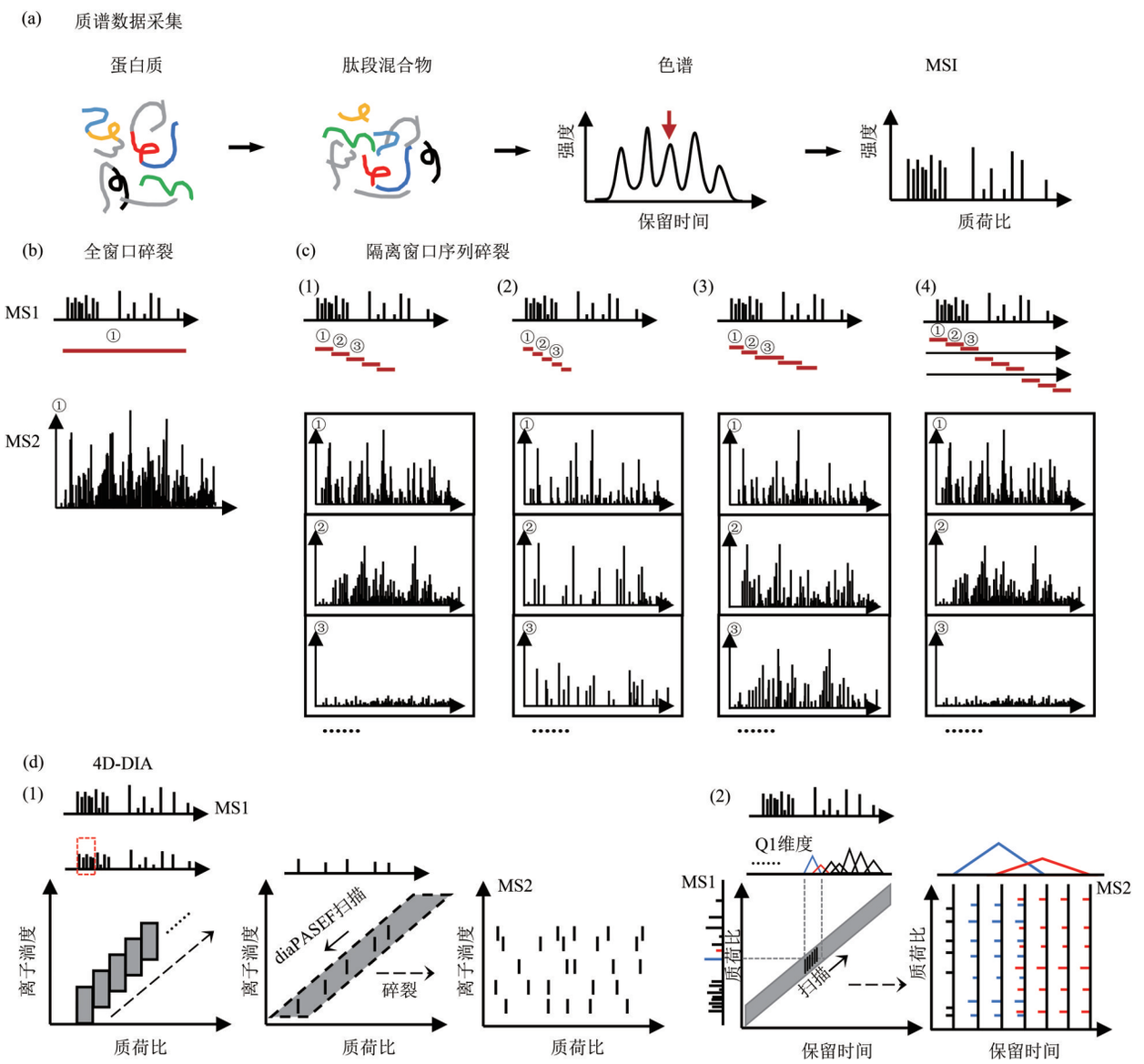


Fig. 2　The workflow of mass spectrometry data acquisition and three different DIA methods

图2　质谱数据采集流程和3种不同的DIA方法

1.1　全窗口碎裂方法

全窗口碎裂方法指每一次循环采集过程中对一级谱图指定较大的质荷比范围（一般大于等于400 u）的母离子共同碎裂，进行一次二级谱图采集，得到的二级谱图包含了所有母离子的碎片离子信息（图2b）。该方法大大提高了质谱仪器的占空比（即质谱仪从离子源采集离子的时间比例）。本节介绍常用的母离子全窗口碎裂方法。

2003年Purvine等[15]提出Shotgun-CID方法，在ESI-TOF-MS质谱仪上分别使用低能量和高能量进行采样，生成两次CID数据。在低能量条件下采集的二级谱图只包含母离子信息，在高能量条件下采集的二级谱图记录了所有肽段的碎片离子信息。作者通过实验证明，Shotgun-CID采集的数据可以通过母离子和碎片离子的色谱曲线来辅助进行肽段鉴定。2006年Plumb等[16]提出了MS^E方法，MS^E在QTOF质谱仪上交替进行低能量和高能量的转换，自动扫描获得母离子、碎片离子信息，获得含碎片离子、母离子和中性损失信息的二级谱图。MS^E方法采集的数据主要用于药物和大分子代谢物的研究。2010年Geiger等[17]提出在Orbitrap Exactive仪器上进行母离子全碎裂（all-ion

fragmentation，AIF）的方法。该方法交替进行母离子和AIF扫描，母离子和AIF扫描各采集1 s，分别采集得到高分辨率的一级和二级谱图。AIF扫描分别使用24、30和36 eV的阶梯式碎裂能量，提高了母离子碎裂效率。

全窗口碎裂方法如Shotgun-CID、MS^E和AIF等高效地采集了较大的指定母离子质荷比范围内的碎片离子信息，有效提高了数据覆盖率和仪器占空比。但是，其生成的二级谱图包含了所有母离子的共碎裂信息，碎片离子干扰几率的增加也对肽段的鉴定造成了困难。因此，母离子全窗口碎裂方法不适用于对复杂样品的大规模分析。

1.2 隔离窗口序列碎裂方法

隔离窗口序列碎裂方法将一级谱图指定母离子质荷比范围划分成多个隔离窗口，依次对每个隔离窗口内的所有母离子碎裂，每次循环采集生成多张二级谱图。相较于母离子全窗口碎裂方法，该方法分别在多个隔离窗口中采集母离子碎裂信息，降低了二级谱图的复杂度。该方法主要可分为4种不同的实现策略，分别为固定大隔离窗口序列碎裂策略、固定小隔离窗口序列碎裂策略、可变隔离窗口序列碎裂策略、优化母离子采集的隔离窗口序列碎裂策略。

1.2.1 固定大隔离窗口序列碎裂策略

固定大隔离窗口序列碎裂策略使用较大隔离窗口（一般大于等于10 u）对母离子质荷比范围进行均匀划分，得到一个隔离窗口序列并依次对其进行碎裂（图2c(1)）。该方法每次循环得到多张二级谱图，所有的二级谱图都有相同的隔离窗口大小。本小节接下来介绍几种常见的采集方法。

2004年Venable等[14]正式提出了DIA名词概念，使用10 u大小的隔离窗口依次遍历母离子质荷比400~1 400的范围，实现对母离子的全面覆盖（为了不产生混淆，本文用原始DIA表示该方法）。作者使用SEQUEST软件并通过扩大候选母离子质荷比范围进行搜索，并利用修改的RelEx软件重构碎片离子色谱曲线。

2012年Gillet等[19]提出了步进式大窗口方法SWATH，采用26 u大小（25 u+1 u重叠部分）的隔离窗口对母离子质荷比400~1 200进行循环采集，每次循环共采集到1张一级谱图和32张二级谱图。在数据解析过程中，Gillet等提出了类似于SRM方法的靶向数据提取策略，并通过mProphet[27]软件进行肽段鉴定。

2019年Amodei等[48]提出了一种大比例重叠窗口DIA方法（overlapping windows DIA），并结合谱图拆分算法来提高母离子的选择性。重叠窗口DIA方法采用窗宽20 u重叠10 u的隔离窗口，交替在质荷比范围500~900和490~890内依次进行隔离窗口扫描。作者利用最小二乘法算法对质荷比范围500~900、490~890的重叠二级谱图进行谱图拆分并去除干扰碎片离子，最终得到10 u大小隔离窗口的二级谱图。

1.2.2 固定小隔离窗口序列碎裂策略

DIA数据解析的有效性与肽段的分离效率直接相关，Heaven等[49]通过实验证明了DIA数据解析的灵敏度与隔离窗口大小为负相关，一系列通过减小隔离窗口的策略也被提出（图2c(2)）。受限于当前质谱仪器的采集效率，难以直接减小隔离窗口的大小，目前该策略主要有两种实现方式，分别为通过相同样品的多次数据采集实现小隔离窗口划分和利用算法拆分大隔离窗口。

第一种策略利用质量区段分离（gas phase fractionation，GPF）技术，将母离子按质荷比划分为多个区间，通过对相同样品多次进样，实现对不同母离子质荷比区间的采集，在不影响循环时间的同时降低了隔离窗口的大小。2009年Panchaud等[18]提出了PAcIFIC方法，将母离子质荷比范围400~1 400划分为67次实验采集，每次实验用10张2.5 u重合1 u大小的隔离窗口实现15 u质荷比范围的覆盖。2021年郭天南团队[50]提出多次采样均匀划分隔离窗口的方法PulseDIA，PulseDIA将传统DIA方法中每个隔离窗口均匀划分并分配到多次气相分离采样中，每次实验对划分的小隔离窗口进行采集，提高了数据灵敏度和数据重现性。小隔离窗口结合多次实验采集策略，有效降低了二级谱图的复杂度，但是增加了质谱采集时间，对实验仪器的稳定性也有较高要求。

第二种策略利用谱图拆分算法对采集到的大隔离窗口二级谱图进行拆分，最终得到多个小隔离窗口的二级谱图。2013年Egertson等[22]提出了多路窗口混合方法MSX。MSX方法对母离子质荷比区间500~900依次划分成100个4 u小隔离窗口，每次随机选取5个小隔离窗口合并碎裂生成二级谱图，直到遍历完所有隔离窗口。MSX方法将采集的二级谱图视为100个4 u小隔离窗口二级谱图叠加得到，通过非负最小二乘法求解得到每个小隔离窗口对应的二级谱图，并能够直接利用较成熟的

DDA数据库搜索软件进行数据解析。

1.2.3　可变隔离窗口序列碎裂策略

上述DIA数据在母离子质量范围上是均匀采集的，但是由于母离子在不同质量的含量是不均匀的，DIA数据在含量高的区域采集到的二级谱图会有更高的碎片离子干扰概率，容易造成二级谱图之间的信息不均等，降低了质谱仪采集效率，同时也对肽段鉴定造成影响。可变隔离窗口序列碎裂策略利用质谱数据的色谱特征和母离子分布情况等特点修改隔离窗口的设置（图2c（3）），进一步减轻了二级谱图的复杂性。设置可变隔离窗口序列有基于算法的自动设置和基于经验的手动设置。

2015年Zhang等[23]提出了以数据为中心（data-centric）的可变母离子隔离窗口DIA方法（variable precursor isolation window DIA），该方法分别实现了根据母离子分布情况（PIP）或总离子流强度（TIC）划分母离子隔离窗口范围的两种采集方式，通过指定隔离窗口数目、质荷比和保留时间范围，自动实现隔离窗口的划分。

2019年Li等[24]提出了随保留时间改变隔离窗口选择的方法RTwinDIA。利用较大质量肽段在反相色谱中通常较晚洗脱的事实，RTwinDIA在较大色谱洗脱时间范围选择更大的母离子质荷比范围，并使用5 u的小隔离窗口依次进行采集。作者评估了一系列DDA搜索软件对RTwinDIA数据的解析能力，结果表明pFind具有直接分析小窗口DIA数据的潜力。事实上，目前大多DIA数据采集方法如SWATH，都会根据经验预设不同大小的隔离窗口进行数据采集。

2020年Guan等[34]提出了DDIA（data dependent−independent acquisition）方法，DDIA结合了DDA方法和DIA方法，在每次循环采集过程中，前0.6 s用于一级谱图和DDA模式二级谱图的采集，后3.0 s采集多张DIA模式二级谱图。该方法利用DDA扫描中鉴定的肽段为DIA扫描的解析提供了先验信息，无需进行额外的DDA数据采集或掺入iRT标准肽段。

1.2.4　优化母离子采集的隔离窗口序列碎裂策略

上述方法都是对二级谱图的采集进行优化，事实上，受限于峰容量限制、母离子干扰和低丰度肽段的离子抑制等影响，一级谱图的母离子信号容易出现干扰或缺失的情况，会对肽段鉴定打分和基于母离子的定量造成影响。为此，一些方法在隔离窗口序列碎裂的基础上优化了母离子的采集。

2014年Thermo团队[51]提出了WiSIM-DIA方法，均匀划分母离子质荷比区间并独立进行一级和二级谱图采集[51]。该方法将母离子质荷比范围400~1 000均匀划分成3个区间，对每个区间的母离子采集一级谱图，并用12 u大小的隔离窗口依次采集二级谱图。独立采集一级谱图的方法提高了母离子的灵敏度和选择性。WiSIM-DIA通过二级谱图进行肽段鉴定，依靠一级谱图母离子定量。

2020年Xuan等[52]提出了基于高分辨率一级谱图定量的采集方法HRMS1-DIA。HRMS1-DIA在母离子质荷比范围400~1 200的循环采集过程中间插入了2张完整母离子质荷比范围的高分辨率一级谱图（图2c（4）），能够检测到更多的母离子信号并重构色谱曲线。在数据分析中，HRMS1-DIA采用了二级谱图定性、一级谱图定量的策略。

除了直接优化母离子采集，还有一些结合BoxCar方法和母离子采集的方法被提出。2021年Mehta等[53]结合BoxCar提出了BoxCarDIA方法，并利用高精度的母离子信息进行肽段鉴定打分和定量。同年Salovska等[54]结合BoxCar和MSX提出了BoxCarmax方法，结合了BoxCar的母离子高灵敏度和MSX的母离子高选择性优点。

多种隔离窗口序列碎裂方法有效地降低了采集到二级谱图的复杂度，有助于实现对DIA数据的深层解析。不过，采集到的二级谱图仍是一系列未知数目母离子的碎片离子信息，难以直接进行肽段鉴定。

1.3　四维DIA数据采集方法（4D−DIA）

上述采集方法得到的二级谱图，破坏了母离子和碎片离子的对应关系，为后续的肽段鉴定造成困难。随着质谱仪器的发展，可以通过获取新的维度信息来重新获得母离子和碎片离子的对应关系。由此引入了离子淌度采集技术和最新的滑动四级杆（sliding quadrupole）技术，采集到额外维度的信息作为传统的只包含质荷比、强度和保留时间的3D-DIA的补充，发展成为4D-DIA数据采集方法。本小节分别介绍基于离子淌度采集技术的DIA-PASEF方法和采用最新滑动四极杆技术的Scanning SWATH方法。

2019年Meier等[25]提出了平行累积连续碎裂（parallel accumulation serial fragmentation，PASEF）的采集方法DIA-PASEF，利用离子淌度质谱仪测量肽段的离子淌度信息来提高母离子选择性。该方法利用捕获离子淌度（trapped ion mobility

spectrometry，TIMS）技术实现离子并行累积，并同步选择四级杆质荷比范围和TIMS设备迁移率范围的母离子，在释放指定淌度的母离子同时使用四级杆进行监测（图2d(1)），极大地提高二级谱图采集效率[55]。DIA-PASEF方法额外采集到的离子淌度信息极大提高了母离子的选择性，有助于后续肽段鉴定，并通过限制离子淌度范围提取到更精准的碎片离子色谱曲线，进一步提高定量精度。作者通过建立含离子淌度的谱库进行靶向数据提取分析DIA-PASEF采集到的数据。在完整蛋白质组消化产物的单次分析中，DIA-PASEF较传统DIA采集多鉴定到了22%的肽段母离子数目。

2021年Markus Ralser团队[26]提出了利用最新滑动四极杆技术的超高速采集方法Scanning SWATH。Scanning SWATH利用四极杆的连续扫描功能，在不损失选择性的前提下拥有更快的循环时间（采集速度接近2 000张/s），结合高流速色谱能够实现5 min甚至0.5 min的高速质谱采集。该方法将四级杆维度的母离子质荷比范围按2 u大小区间进行划分，所有重叠于区间范围内的碎片离子强度被加和到对应区间中（图2d(2)）。随着滑动窗口逐渐通过母离子质荷比，对应碎片离子信号先出现后消失，累计到区间中的强度为三角形状的剖面，最高信号为母离子质荷比对应的区间。通过比较碎片离子在Q1四级杆维度上的强度变化，能够分配母离子质荷比来提高母离子选择性。

4D-DIA方法如DIA-PASEF和Scanning SWATH，通过记录了额外的离子淌度和四级杆维度母离子信息，在一定程度上重构了二级谱图中母离子和碎片离子的关系，进一步提高了数据解析能力。

综上所述，DIA数据采集方法如全窗口碎裂方法、隔离窗口序列碎裂方法和4D-DIA方法都有效实现了蛋白质样品的高通量采集，主要区别在于采集到二级谱图的复杂程度，目前最常用的数据采集方法是SWATH或可变窗口SWATH以及DIA-PASEF方法。全窗口碎裂方法采集到的二级谱图包含了全窗口范围内的母离子，谱图解析的复杂度较大。隔离窗口序列碎裂方法通过多种采集策略减少了二级谱图的母离子数目和隔离窗口大小，有效降低了谱图解析的复杂度。随着质谱仪器的发展，DIA采集二级谱图隔离窗口大小有可能接近于DDA二级谱图隔离窗口大小，实现DIA和DDA解析流程的融合。4D-DIA方法通过额外采集的数据维度获取母离子和碎片离子的对应关系，提高了母离子的选择性，大大降低了谱图解析的复杂度。4D-DIA方法也是未来DIA数据采集的重要发展方向。

2 DIA数据解析

DIA数据解析是指对DIA方法采集到的质谱数据进行肽段鉴定，得到肽谱匹配。由于肽段在DIA数据中连续采集了多张二级谱图，本文的肽谱匹配特指一条肽段和单张或多张连续二级谱图的匹配情况。传统DDA搜索软件难以直接解析DIA数据的二级谱图，因此需要专门针对DIA数据的搜索算法。本节首先解释传统DDA搜索方法解析DIA数据二级谱图的难点，包括母离子质荷比难以确定和碎片离子难以区分两大难点。然后介绍目前常用的DIA数据解析方法，根据不同的搜索策略可分为谱库搜索方法、蛋白质序列库直接搜索方法、伪二级谱图鉴定方法和从头测序4种方法（图3）。

2.1 DIA二级谱图数据解析的难点

DIA二级谱图数据解析方法通过匹配二级谱图的肽段碎片离子信息，实现肽段的鉴定。如何解析混合二级谱图是实现DIA数据解析的关键技术难点。由于DIA方法对隔离窗口范围内的所有母离子进行碎裂，母离子和碎片离子的对应关系被打破（图4），对DIA二级谱图的解析造成困难。

2.1.1 母离子质荷比难以确定

母离子质荷比的确定是DIA二级谱图数据解析的一大挑战。传统DDA软件的二级谱图解析算法通过选择较小母离子质量误差范围内的候选肽段缩小搜索空间，其搜索空间和母离子碎裂的隔离窗口大小呈正相关，母离子质荷比是否确定影响了搜索效率和鉴定灵敏度。对于DIA二级谱图，由于所包含的肽段母离子的质荷比无法确定，DDA软件无法通过母离子质量获取候选肽段，同时较大的隔离窗口无法有效缩小搜索空间，增加的候选肽段数目提高了搜索所需的时间，也导致了更高的鉴定假阳性率。

2.1.2 碎片离子难以区分

DIA二级谱图数据解析的另一挑战是难以区分多个肽段共碎裂生成的碎片离子。在二级谱图解析过程中，针对二级谱图实际包含的不同肽段数目，一般采用不同的搜索策略和打分公式，如传统DDA软件通常将二级谱图视为单个肽段碎裂生成，通过设计有效的单肽打分函数进行肽谱匹配。而来

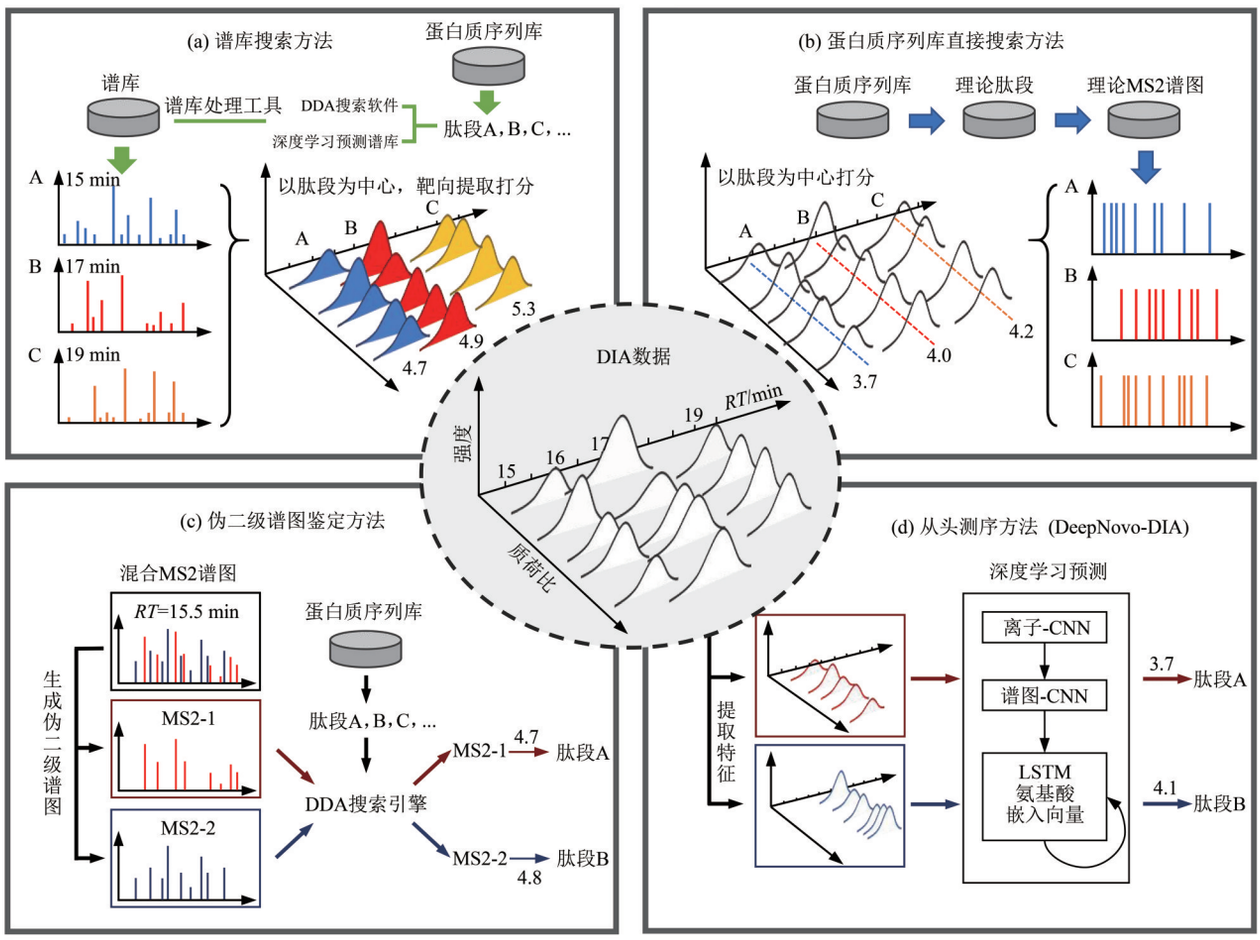


Fig. 3　The workflow of DIA data analysis

图3　DIA数据解析流程

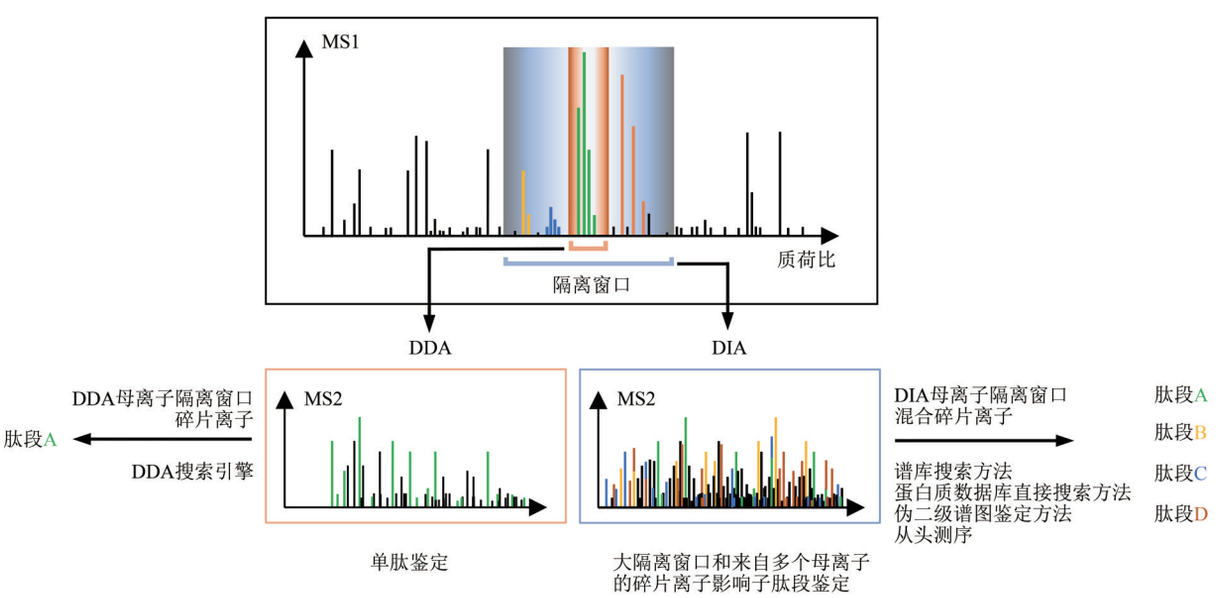


Fig. 4　Difficulties in identifying DIA MS/MS spectra

图4　鉴定DIA二级谱图的难点

自于DIA二级谱图的多个肽段的碎片离子和相同质荷比的干扰离子，对单个肽段的匹配打分造成了干扰，影响其鉴定精度。理论上，多个肽段的组合随着共碎裂肽段数目呈指数级别增加，其搜索空间大大增加。

通过上述两点可知，传统的DDA搜索软件难以实现对DIA数据大规模的可靠解析。因此，需要针对DIA的数据解析方法。为了解决从混合二级谱图中解析得到多个肽段的问题，发展出了4种不同的搜索方法：第一种方法是通过已知肽段鉴定结果的谱库去匹配DIA二级谱图实现肽段鉴定；第二种方法是基于蛋白质序列库对肽段理论二级谱图和实际二级谱图进行匹配；第三种方法是对DIA二级谱图解卷积拆分成伪二级谱图，再结合DDA搜索软件进行搜索；第四种方法是不利用已知序列库和谱库信息，直接对DIA二级谱图进行从头测序。接下来分别介绍这4种方法。

2.2 谱库搜索方法

谱库搜索方法利用已知肽段鉴定信息的谱库，和实际二级谱图进行匹配以实现肽段鉴定，是目前最常用的DIA数据解析方法（表2）。谱库搜索方法的概念最早由Yates等[56]提出，后经Lam等[57-59]实现了完整的谱库构建和搜索流程用于DDA数据分析和SRM、PRM数据的靶向分析，2012年该方法被Aebersold团队[19]引入到SWATH数据中进行靶向分析，并逐渐广泛用于各种DIA数据的分析。

谱库搜索方法的流程主要包括两部分，分别为谱库构建和数据提取。首先根据DDA数据的鉴定结果或深度学习预测结果生成谱库，然后对谱库中每个肽段进行数据提取，提取母离子、碎片离子色谱曲线或二级谱图匹配等特征进行肽段打分，最终得到肽谱匹配（图3a）。

2.2.1 谱库构建

谱库是一系列二级谱图及对应肽段鉴定结果的集合，包括每个肽段的母离子和碎片离子的质荷比、电荷和强度以及保留时间信息。目前的谱库构建方法可分为DDA数据鉴定结果建库和深度学习预测建库。此外，还有一些工作对谱库构建进行了优化。

DDA数据鉴定结果建库方法利用DDA数据鉴定到肽段的二级谱图和保留时间等信息生成谱库。DDA鉴定结果直接影响了谱库的规模和质量。常用的DDA搜索软件有pFind、MSFragger和SEQUEST以及MaxQuant等。使用与DIA实验相同条件的DDA数据能提高鉴定的准确度，但是其结果局限于DDA软件的鉴定结果，同时失去低丰度肽段检测的优势。而且，采集DDA数据生成谱库的方法有较高制作成本，而使用已有的大规模谱库也会因为实验环境、仪器和样品之间的差异降低谱库搜索的质量[60]，提高鉴定数目的同时降低了鉴定准确度[61-62]。

近年来深度学习在蛋白质组学中有着广泛的应用[63-65]，为了摆脱建库时对DDA实验谱图的依赖，提出了深度学习预测谱库方法。2017年Zhou等[66]首次提出了基于深度学习的谱图预测工具pDeep，利用双向长短时记忆网络（Bi-LSTM）对肽段理论二级谱图进行预测。2019年Gessulat等[67]提出了谱库预测算法Prosit，使用双向门控循环单元（Bi-GRU）网络模型并结合注意力机制对理论二级谱图和保留时间进行预测。2020年乔亮团队[68]进一步提出DeepDIA算法，结合CNN和LSTM来预测谱库，并使用单独的深度网络模型来预测肽段的可检测性。作者还发现，相较于物种变化，实验仪器的变化对预测准确性影响更大。深度学习预测谱库方法能够生成整个蛋白质序列库的完整谱库，极大地提高了鉴定深度，同时为准确鉴定低丰度肽段提供了可能。

DDA数据鉴定或深度学习预测得到的谱图，通常需要使用专门软件如EasyPQP、SpectraST[58]和Skyline[69]等生成谱库。谱库的生成质量直接影响了后续鉴定结果的灵敏度和准确度，为此也提出了一些优化谱库结果的软件。Midha等[70]提出了谱库质量控制软件DIALib-QC，通过62个一致性参数来评估谱库的完备性和准确性，进一步提供优化选项。Zhu等[71]提出了谱库构建自动化流程，该流程支持pFind的鉴定结果，通过基于Docker的服务器替代本地运行并结合SpectraST构建谱库。

2.2.2 基于碎片离子特征提取的方法

基于碎片离子特征提取的方法获取谱库中每个肽段的保留时间和前n个强度的碎片离子，以肽段为中心去DIA数据中进行靶向信息提取，通常根据保留时间指定一段时间范围，并根据母离子和碎片离子质荷比提取色谱曲线等特征进行匹配打分，实现肽谱匹配。该方法对每个肽段判断数据中能否检测到该肽段，而不是直接考虑二级谱图对应的肽段[72]，无需考虑二级谱图中来自其他肽段的碎片离子，在缩小搜索空间的同时提高了肽段鉴定灵

Table 2　Commonly used software tools for spectral library search method

表2　常用谱库搜索方法软件相关信息

软件名称	年份	数据库类型	描述	重排序方法	软件下载
mProphet	2011	谱库	靶向提取SRM/DIA数据的色谱曲线信息，使用色谱曲线共洗脱、相似度和标记肽段相关性等特征进行打分	LDA半监督学习	http://www.mprophet.org/
OpenSWATH	2014	谱库	对提取到的谱峰组计算包含基于色谱的打分、基于library的打分、基于二级谱图的多项特征用于打分和重排序	PyProphet	http://www.openswath.org
Spectronaut	2014	谱库/蛋白质序列库	重写了mProphet，并引入了iRT对保留时间进行校正来提高鉴定精度	Pyprophet	https://biognosys.com/software/spectronaut/
DIA-NA	2015	谱库	基于碎片离子之间的期望比值计算色谱峰的打分，mProphet优化为PyProphet进行重排序	PyProphet	http://quantitativeproteomics.org/diana
MSPLIT-DIA	2015	谱库	以谱图为中心进行非靶向谱库搜索，对谱库肽段的碎片离子和实际谱图计算点击打分	无	http://proteomics.ucsd.edu/software-tools/msplit-dia/
SWATHProphet	2015	谱库	两阶段打分：首先校准谱库保留时间，然后根据保留时间范围靶向数据提取打分	iProphet	http://tools.proteomecenter.org/software/SWATHProphet/
Skyline	2017	谱库	通过CRAWAD算法计算一阶导数和二阶导数得到谱峰极值和拐点定义谱峰组	mProphet	https://skyline.ms/
EncyclopeDIA	2018	谱库/色谱库	谱库/色谱库靶向提取，在小隔离窗口鉴定生成添加色谱形状的色谱库，提高鉴定精度	Percolator 3.1	https://bitbucket.org/searleb/encyclopedia/wiki/Home
Specter	2018	谱库	利用非负最小二乘法计算得到各肽段的强度系数，对二级谱图进行碎片离子和谱峰强度的拆分	LDA模型	https://github.com/rpeckner-broad/specter
Prosit	2019	蛋白质序列库	使用Bi-GRU结合注意力机制对输入的embedding肽段序列进行编码解码预测二级谱图和保留时间	无	http://www.proteomicsdb.or/ prosit
DIA-NN	2019	谱库/蛋白质序列库	对靶向提取的谱峰组得到最优碎片离子并根据色谱曲线相关性计算得到72维特征用于谱峰的选择和重排序	深度神经网络	https://github.com/vdemichev/DiaNN
Deep-DIA	2020	蛋白质序列库	使用CNN+LSTM的网络结构对One-hot表示的肽段序列进行谱库预测	无	https://github.com/lmsac/DeepDIA/
MaxDIA	2021	谱库	通过多次迭代搜索不断限制搜索空间和提高校准质量，实现肽段鉴定结果的优化	XGBoost半监督学习算法	https://www.maxquant.org/maxquant/
DreamDIA	2021	谱库	使用预训练深度学习模型对多种类型的离子色谱曲线提取高维特征，实现肽段鉴定	XGBoost半监督学习	https://github.com/xmuyulab/DreamDIA-XMBD

敏度。

2011年Reiter等[27]开发了自动处理和评估SRM数据的软件mProphet，后被Gillet等[19]用于SWATH数据解析。mProphet对每个肽段通过靶向数据提取得到多组候选谱峰组（肽段母离子、碎片离子在质谱数据中的谱峰信号），并对谱峰组计算色谱曲线共洗脱和相似度、理论和实际碎片离子强度相关性和轻重标记肽段相关性等多个打分，用于后续重排序并根据判别分数计算最优谱峰组。mProphet在分析中引入诱饵库策略，通过对目标库肽段反转或质量偏移生成诱饵库，合并搜索并进行可信度评估。

2014年Rost等[28]发表了第一款全自动流程的DIA谱库搜索软件OpenSWATH，鉴定流程分为数据转换、保留时间校正、色谱曲线提取、谱峰组打分、统计分析5个部分。OpenSWATH引入索引保留时间（indexed retention time，iRT）进行保留时间校正，对每个标准肽段搜索其在质谱数据所有保留时间范围内的最佳谱峰组，将最佳峰组的保留时间与相应iRT值进行拟合，通过拟合函数对剩余谱库的保留时间进行校正。然后OpenSWATH对提取到的谱峰组进行打分，考虑色谱曲线、谱库与实际谱图相关性和色谱曲线峰值处的肽谱匹配打分等多项特征。由于OpenSWATH便捷的特点，QuantPipe[72]和DIAProteomics[73]移植了OpenSWATH并结合质控算法PyProphet[30]实现了完整的鉴定流程，GproDIA也通过调用OpenSWATH来帮助实现糖肽的鉴定[74]。MacCoss团队[75]提出的Skyline近期版本也支持了对DIA数据的谱库搜索，通过色谱曲线检测算法CRAWDAD[76]识别谱峰范围，进行肽段鉴定。

2015年Johan Teleman等[30]提出了DIANA算法，对谱库中每个肽段计算母离子、碎片离子之间的马尔科夫比值概率和Pearson相关系数，并作为特征使用PyProphet进行重排序。DIANA通过比较谱库中谱峰和DIA谱图提取谱峰强度的一致性来识别干扰碎片离子，并根据谱库母离子、碎片离子强度比值一致性对其色谱曲线的面积进行校正。同年Keller等[31]提出了SWATHProphet，其鉴定流程主要包括保留时间校正和谱峰组打分。SWATHProphet提出了两种方法筛选干扰碎片离子，第一种方法通过计算相近保留时间内肽段的共碎裂情况来识别谱库中其他肽段的干扰碎片离子，第二种方法根据谱库中碎片离子相对强度计算肽段各碎片离子色谱曲线的相关性来筛选谱库外的肽段或噪声的干扰离子。

2018年MacCoss等[32]发表的EncyclopeDIA软件提出了构建色谱库进行靶向分析的方法。EncyclopeDIA通过DDA搜索软件或者Walnut（对PECAN[39]进行重写的方法）搜索多次GPF采样的DIA数据，对得到的鉴定结果生成较谱库多了碎片离子色谱信息的色谱库。EncyclopeDIA使用X! Tandem的加权点积公式对肽段保留时间点上的谱图和色谱库进行打分，计算保留时间上最高分对应的时间点的谱图匹配打分、碎片离子准确度打分和母离子准确度打分等15个辅助匹配特征进行最终肽段鉴定。新版本EncyclopeDIA也支持了使用Prosit算法直接预测生成谱库。

2019年Vadim Demichev等[33]提出了DIA-NN软件，通过深度神经网络进行肽段鉴定，实现对DIA数据的高通量蛋白质覆盖。DIA-NN支持DDA建库和Prosit算法预测建库。DIA-NN对每个肽段结果靶向获取谱峰组，对每个谱峰组，依据前6个最高强度碎片离子的色谱曲线相似度，找到代表肽段整体色谱曲线变化情况的最优碎片离子，并据其计算得到73维特征分别用于最优谱峰组的选择和肽段最终打分。新版的DIA-NN（1.8版本）通过修改打分细节和添加新的特征，还支持了对DIA-PASEF[77]和Scanning SWATH[26]数据的分析。

2020年Guan等[34]提出了DDIA，其分析流程包括DDA鉴定、校准保留时间曲线、训练DIA提取分类器和DIA数据提取4个部分。首先，使用MSGF+软件对DDA扫描采集的二级谱图进行肽段鉴定，利用肽段鉴定结果进行保留时间的校准和DIA提取分类器的训练，并使用Skyline实现DIA扫描的肽段鉴定。利用肽段在DDA和DIA扫描中具有相同保留时间的原理，对DDA扫描鉴定到的保留时间和深度学习预测的保留时间进行校正，获得校准曲线并对谱库所有结果的保留时间完成校正，实现后续谱库搜索。

2021年Pavel Sinitcyn等[35]提出了MaxDIA软件，MaxDIA软件支持DDA建库（MaxQuant搜索）和深度学习预测谱库（基于双向循环神经网络的DeepMass: Prism算法[78]）两种方法，通过多次迭代搜索优化肽段鉴定结果。MaxDIA首先按照预设置的母离子、碎片离子质量误差去谱库记录的一定保留时间范围内进行搜索，根据搜索结果建立线性保留时间校正函数和母离子、碎片离子校正函数，限制其搜索空间并进行迭代搜索。MaxDIA实现了深度蛋白质组覆盖，并通过将MaxDIA与两项新技术（BoxCar采集和TIMS技术）相结合，实现了对蛋白质组的深度与精确定量。

2021年韩家淮和俞容山团队[36]提出了基于深度学习的鉴定软件DreamDIA。不同于其他DIA软件使用启发性特征，DreamDIA使用预训练LSTM模型对多种离子的色谱曲线提取高维特征，实现肽段鉴定。DreamDIA首先随机选取部分谱库结果进行保留时间归一化，通过LSTM模型对肽段结果进行全局保留时间范围的遍历打分，确定最佳保留时间并和谱库的保留时间进行拟合，从而预测剩余肽

段的保留时间。然后，DreamDIA对所有结果提取基于谱库碎片离子、母离子、未碎裂母离子和对应的同位素峰等共170个多种离子类别的色谱曲线矩阵，结合LSTM模型和全连接网络进行打分，得到最优匹配结果。

2.2.3 基于二级谱图匹配的方法

基于二级谱图匹配的方法将二级谱图视为多个肽段共碎裂生成的混合谱图，以谱图为中心进行谱库搜索。2015年Wang等[79]提出了MSPLIT-DIA，根据二级谱图的母离子质量误差筛选候选谱库，对每个谱库结果获取各谱峰在二级谱图中质量误差（50 ppm）范围内的最高谱峰，并计算和二级谱图的余弦相似度打分。为了避免多个干扰碎片离子造成错误匹配，MSPLIT-DIA对相似度大于0.7的谱库只保留一个打分最高的结果，对匹配的碎片离子提取色谱曲线进行相似度打分。MSPLIT-DIA还支持对谱库进行保留时间校正来提高鉴定率。2018年Peckner等[43]提出了Specter软件，该方法假设二级谱图由母离子共碎裂的碎片离子线性叠加生成，根据谱库中记录的碎片离子相对强度信息，利用非负最小二乘法将二级谱图拆分成谱库中多个肽段的线性组合，得到肽段结果对应的加权系数。MSPLIT-DIA利用拆分后的碎片离子进行定量，在一定程度上去除了碎片离子干扰。

综上所述，谱库搜索方法特别是以肽段为中心的靶向数据提取方法没有直接对DIA混合二级谱图进行解析，而是根据谱库的保留时间和碎片离子相对强度信息进行靶向数据提取，避免了二级谱图共洗脱肽段之间的干扰情况，提高了鉴定灵敏度，同时有更好的定量精度。谱库搜索方法较蛋白质序列库搜索方法具有更高的灵敏度[80]和数据重现性[81]，对应的搜索速度也更快，部分原因是谱库的搜索空间较蛋白质序列库的搜索空间更小。理论上，常用谱库的肽段数目少于蛋白质序列库理论酶切后的肽段数目，极限条件下使用深度学习预测蛋白质序列库得到谱库的规模和蛋白质序列库相当。

谱库搜索方法也存在一些不足和优化方向。首先，该方法对谱库的完备性和正确性有较高要求。一方面，重复采集DDA数据的谱库生成方法具有高灵敏度，但其肽谱匹配数目也受限于DDA鉴定结果，难以将低丰度肽段鉴定出来并放到谱库。另一方面，基于深度学习的预测建库方法虽然增加了谱库中的肽谱匹配数目，但也增加了随机匹配概率，同时在不同仪器和物种上的灵敏度仍需进一步提高。其次，该方法目前没有考虑开放式搜索[82]，直接对母离子隔离窗口内的二级谱图进行匹配，无法实现对意外修饰肽段的鉴定。结合谱库信息和开放式搜索技术，能够在提高DIA数据解析率的同时对蛋白质样品的生物意义进行更深刻地挖掘，利用谱库进行DIA数据的开放式搜索是实现谱图深度解析的一个发展方向。

2.3 蛋白质序列库直接搜索方法

蛋白质序列库直接搜索方法无需事先构建谱库或拆分混合二级谱图，和传统DDA搜索类似，首先将序列库内蛋白质理论酶切生成肽段，而后直接对肽段理论二级谱图和DIA实验二级谱图进行匹配（图3b）。根据搜索策略的不同，该方法可分为以谱图为中心策略和以肽段为中心策略。表3列举了该方法常用的软件及相关信息。

2.3.1 以谱图为中心的搜索策略

以谱图为中心的搜索策略对DIA二级谱图直接解析，即利用传统DDA软件的肽谱匹配打分计算二级谱图对应的多个肽段。

在早期，人们尝试利用传统的DDA搜索策略直接解析DIA数据。Venable等[14]在原始DIA数据中使用SEQUEST软件[83]对10 u大小隔离窗口的二级谱图进行单肽鉴定。Li等[24]在RTwinDIA数据中使用pFind对5 u大小隔离窗口的二级谱图进行搜索。随着仪器采集效率的提高，使用DDA软件直接鉴定小窗口DIA数据逐渐变成可能。

DDA软件通常需要指定母离子质荷比，扩大候选母离子范围搜索DIA数据的方法容易导致更多假阳性结果，为此，一些方法通过指定二级谱图的母离子质荷比进行谱图解析。2006年Venable等[84]在计算二级谱图对应母离子质荷比的方法中提到该方法能用于DIA数据解析。Aebersold团队[85]提出的ProbIDtree用于解决DDA数据的混合谱图鉴定问题，通过迭代剪枝去除二级谱图已匹配的谱峰实现肽段鉴定，具有解析DIA数据的潜力。

一级谱图上母离子信号干扰和较大隔离窗口范围阻碍了以谱图为中心的搜索方法，为此提出了直接对二级谱图进行肽段鉴定的方法。2021年Lu等[37]提出了DIAmeter软件，利用二分图进行肽谱匹配和候选肽段的过滤。DIAmeter将二级谱图集合和模拟酶切肽段集合视作二分图中两个互不相交的子集，将肽谱匹配表示为两个节点的边。首先，DIAmeter使用XCorr打分函数对每张二级谱图和隔离窗口内1~5电荷的母离子打分并初步筛选得到候

Table 3 Commonly used software tools for protein database search method

表3 常用蛋白质序列库直接搜索方法软件相关信息

软件名称	年份	描述	搜索策略	软件下载
ProbIDtree	2005	以原始二级谱图为根节点，用边表示各候选母离子的概率，减去匹配肽段的谱峰后的二级谱图作为子节点，迭代进行肽段匹配	谱图中心	N/A
QCorr	2006	利用b/y互补离子计算二级谱图多个母离子质量，用DDA搜索引擎分别进行打分	谱图中心	http://fields.scripps.edu/download/QCorr.ZIP
FT-ARM	2012	以肽段为中心，对每条肽段在色谱时间上计算理论谱图和实际谱图的点积打分，得到打分色谱	肽段中心	http://brucelab.gs.washington.edu/
PECAN	2017	引入背景蛋白质去除背景干扰，以肽段为中心对肽段碎片离子色谱曲线进行打分，利用背景库的打分去除干扰	肽段中心	http://pecan.maccosslab.org
pFind3.0	2017	DDA搜索引擎，支持开放式搜索，在RTWinDIA采集方法中搜索5 Da大小隔离窗口的DIA二级谱图	谱图中心	https://github.com/pFindStudio/pFind3/
DIAmeter	2021	将二级谱图和蛋白质序列库视作二分图，使用Xcorr打分得到初步候选母离子，计算一系列反映匹配情况的特征对候选母离子进行进一步过滤	谱图中心	http://crux.ms

选母离子，然后对每条边计算包括修正XCorr打分[86]、母离子强度、碎片离子匹配打分、预测保留时间差值以及母离子和碎片离子色谱曲线相关性共5个特征，对特征线性组合打分用于进一步母离子筛选，通过不断对二分图进行边的插入和删除实现肽段的鉴定。

该策略提供了直接鉴定低丰度肽段和未知修饰肽段的可能，但是来自不同肽段的碎片离子和大量干扰碎片离子会对肽谱匹配打分造成影响，一定程度上降低了鉴定结果的可信度。如何更精确地确定肽段质荷比，设计更有效的结合保留时间维度信息和单张谱图匹配信息的打分是提高鉴定灵敏度的优化方向。

2.3.2 以肽段为中心的搜索策略

另一种以肽段为中心的搜索策略，不是为每张二级谱图分配一个或多个最优的肽段，而是将肽段和多个二级谱图做点积运算并结合启发式搜索进行打分，报告每个肽段对应的最佳匹配结果。该策略避免了对二级谱图的多个母离子进行拆分[87]。

2012年Weisbrod等[38]提出了针对大窗口数据的解析算法FT-ARM。FT-ARM用于100 u的大隔离窗口数据，考虑肽段2+和3+电荷的母离子并通过SSRCalc算法预测肽段保留时间。对保留时间范围的二级谱图序列，依次计算理论谱图和实际谱图的点积得到肽段在保留时间范围上的打分列表，并将打分除以谱图上碎片离子个数来提高信噪比。FT-ARM不依赖一级谱图的母离子信息，能够鉴定到一级谱图中没有信号强度的肽段。

2016年MacCoss团队[39]根据FT-ARM的思想提出PECAN，通过引入背景数据库帮助进行肽段鉴定。对每条感兴趣肽段，PECAN使用背景数据库的碎片离子频率倒数作为碎片离子谱峰权重，生成肽段的理论谱图。然后，PECAN对理论谱图向量和提取的碎片离子色谱曲线矩阵进行打分，通过减去诱饵背景数据库的平均打分去除干扰，得到打分向量并报告最优的打分和保留时间。此外，PECAN还可以直接通过小隔离窗口DIA数据生成谱库，用于后续大窗口数据的分析。

以肽段为中心的搜索策略利用了碎片离子在二级谱图上连续的特点，对肽段在色谱保留时间上的连续多个打分进行匹配，提高了鉴定灵敏度。但该策略对每张二级谱图和肽段打分的方式增加了计算复杂度，同时点积打分没有考虑相对离子强度和碎片离子干扰问题，其鉴定灵敏度不如谱库搜索方法。

2.4 伪二级谱图鉴定方法

伪二级谱图方法不直接解析DIA原始二级谱图，而是利用谱图拆分算法将二级谱图进行拆分得到多张包含单个肽段碎片离子的伪二级谱图，再结合传统DDA软件搜索伪二级谱图（表4）。该方法通过对二级谱图进行预处理，重建母离子和碎片离子的对应关系，降低了谱图复杂度和谱图解析难度（图3c）。本节介绍常用的伪二级谱图鉴定软件。

2009年Bern等[40]提出了基于碎片离子色谱矩阵聚类的DeMux软件。DeMux首先对各隔离窗口采集的1 440张二级谱图集合向量化得到1 440×1 200（质荷比范围0~1 200）的碎片离子色谱矩阵，并按保留时间划分得到多个100×1 200的小矩

Table 4　Commonly used software tools for pseudo−MS/MS spectra identification method

表4　常用伪二级谱图鉴定方法软件相关信息

软件名称	年份	是否使用互补离子	描述	重排序方法	软件下载
DeMux	2010	否	将二级谱图序列转化为二维矩阵，按碎片离子强度相关系数进行聚类生成伪二级谱，结合Byonic鉴定	Byonic内置重排序	N/A
DIA-Umpire	2015	是	利用色谱曲线相关性获取母离子-碎片离子对，输出满足阈值和保留时间差值的碎片离子生成伪二级谱图	DDA软件内置重排序	http://diaumpire.sourceforge.net/
Group-DIA	2015	否	利用母离子-碎片离子色谱曲线在多个数据中的一致性来筛选生成二级谱图，使用MasCot搜索	MasCot内置重排序	http://yuanyueli.github.io/group-dia/
CorrDec	2020	否	计算碎片离子对所有候选母离子的皮尔逊相关系数，通过打分标准实现母离子的分配和噪声峰的删除	/	http://prime.psc.riken.jp/compms/msdial/main.html

/：文献中没有介绍。

阵。然后对小矩阵按列强度排序，根据相似度按列聚类得到多个肽段对应的碎片离子色谱曲线簇c，对每个肽段对应的特征c按列求和得到色谱曲线Elute（c）以及按行筛选各列碎片离子生成伪二级谱Synth（c）。最后，使用Byonic对伪二级谱图进行肽段鉴定，使用得到的色谱曲线分别计算强度。

2015年Tsou等[41]提出基于母离子-碎片离子共洗脱进行二级谱图拆分的方法DIA-Umpire。DIA-Umpire获取一级谱图上母离子、二级谱图上未碎裂母离子和碎片离子的色谱曲线，根据皮尔逊相关系数和保留时间差值计算母离子-碎片离子的匹配情况。对匹配到的母离子和碎片离子峰簇，输出互补碎片离子和满足一定相关系数、保留时间差值的碎片离子，最终生成伪二级谱图。DIA-Umpire可以使用X！Tandem[88]、Comet[89]和MSGF+[90]搜索软件对伪二级谱图进行序列库搜索。

同年韩家淮团队[42]提出了Group-DIA软件，利用肽段在多个数据中色谱曲线的一致性来确定母离子-碎片离子对。Group-DIA首先利用保留时间校正算法ChromAlign[91]对齐多个数据的肽段保留时间，通过肽段的母离子-碎片离子在多个数据中的相对强度分布一致性的假设，合并肽段在多个数据中的母离子和碎片离子色谱曲线，比较其在所有数据之间的相关性并筛除不属于该肽段的碎片离子。Group-DIA通过传统DDA序列库搜索软件如Mascot实现肽段鉴定，在多数据分析中相较DIA-Umpire鉴定到更多的肽段数目和更多的低丰度肽段。

2020年Tada等[44]提出了利用反卷积方法的代谢组DIA分析工具CorrDec，基于母离子和碎片离子之间的谱峰强度在多个数据之间一致性的假设拆分二级谱图。CorrDec以二级谱图为单位对多个数据的相应二级谱图去卷积，通过相似度计算实现对二级谱图谱峰的母离子分配，得到所有碎片离子对每个母离子的相似度打分，然后利用打分标准去除噪声和干扰碎片离子，生成伪二级谱图。该方法目前已被整合到代谢组分析平台MS-DIAL[92]。

伪二级谱图方法主要利用色谱曲线一致性来生成伪二级谱图，并利用DDA软件进行搜索，能够对DIA数据进行深层解析如开放式搜索。不过该方法也有局限性，如DIA-Umpire无法拆分没有母离子信号的肽段。而且，受限于二级谱图的离子干扰、离子抑制情况和谱图拆分算法，伪二级谱图方法拆分得到的二级谱图数目偏少，导致其鉴定数目较谱库搜索方法较少。未来，结合深度学习算法识别母离子、碎片离子色谱曲线的高维表征，并利用4D-DIA数据提供的更高的母离子选择性，能够进一步提高二级谱图拆分能力。

2.5　从头测序方法

上述3种方法利用谱库或蛋白质序列库对DIA数据进行解析，其报告的肽段结果局限在所用的数据库中。从头测序方法不对肽段序列做任何限制，直接依靠二级谱图推断肽段序列。DIA数据包含碎片离子在保留时间维度上的信息，可以通过对多张连续二级谱图的匹配进行肽段鉴定。目前，DIA数

据从头测序使用的方法主要是深度学习预测方法。

2019 年 Tran 等[45] 发表了 DeepNovo-DIA 方法，通过 DIA-Umpire 利用色谱曲线一致性检测母离子和碎片离子特征，从 DIA 数据中提取肽段保留时间范围内的多张伪二级谱图，再利用卷积网络和 LSTM 模型捕捉三维数据之间的相关性并考虑肽段序列模式，实现从头测序（图 3d）。值得注意的是，DeepNovo-DIA 通过打分阈值过滤测序结果，没有对结果进行可信度评估。

DeepNovo-DIA 具有鉴定到未知物种肽序列的优点，但是因其需要进行谱图拆分，损失了肽段在原始二级谱图中的部分谱峰，同时没有可信度评估，鉴定结果灵敏度和精确度不如谱库搜索方法。未来，直接对 DIA 原始二级谱图使用深度学习实现从头测序，并考虑对结果进行可靠性评估，将加速推动 DIA 从头测序方法的实用化。

3 肽段鉴定结果重排序与假发现率估计

不管是 DDA 数据分析还是 DIA 数据分析，由于搜索软件在肽谱匹配过程中可能出现错误匹配（又称随机匹配），同时缺少对肽谱匹配的可信度评估，所以不能直接将软件报告的所有结果用于后续蛋白质推断和定量分析，需要根据可信度水平对肽段鉴定结果进行重排序，并对报告的肽谱匹配集合进行假发现率估计。DIA 肽段鉴定结果重排序与假发现率估计的原理与 DDA 基本相同，只是在具体的实现细节上有所不同。

3.1 肽段鉴定结果重排序

由于各 DIA 数据解析方法的搜索策略和打分方式不同，各方法所用的 DIA 重排序算法也有所不同，本节对各 DIA 数据解析方法使用的 DIA 重排序算法进行介绍。

3.1.1 从头测序方法的重排序

现有的从头测序方法由于没有已知的蛋白质数据库或谱库，没有构建诱饵结果，直接使用经验打分阈值过滤出正确结果，无法对鉴定结果进行可信度评估和重排序。

3.1.2 伪二级谱图鉴定方法的重排序

伪二级谱图鉴定方法的重排序算法一般由搜索过程中使用的 DDA 软件实现。Percolator 是 DDA 软件最广泛使用的重排序算法[93]。该算法将重排序看成一个二分类问题，即区分正确肽谱匹配（正例）和错误肽谱匹配（负例）。正负例的选择通过构建诱饵库实现，通常使用诱饵库的匹配结果作为负例。该算法对每个肽谱匹配计算 20 维特征向量，然后进行多次迭代学习，每次迭代过程选择对应目标库的高可信肽谱匹配作为正例，选择对应诱饵库的肽谱匹配作为负例，训练支持向量机（support vector machines，SVM）模型。通过训练好的模型对所有肽谱匹配进行重打分，得分高的为正例，得分低的为负例。

3.1.3 蛋白质序列库直接搜索方法的重排序

蛋白质序列库直接搜索方法的重排序算法与其使用的搜索策略相关。以谱图为中心的搜索策略使用 DDA 软件实现肽谱匹配，对应的重排序算法也通过 DDA 搜索软件内置的重排序算法实现，选取的特征主要根据单张二级谱图匹配信息计算得到，如肽谱匹配打分、母离子质量误差、碎片离子质量误差等，最后结合机器学习算法实现重排序。以肽段为中心的搜索策略将肽段和保留时间范围内的多张二级谱图进行匹配，对应的重排序算法不只考虑单张谱图，而是考虑一条肽段对应的相近保留时间范围内的所有谱图的匹配，即以肽段为单位进行重排序。如 PECAN 对匹配结果中目标库和诱饵库的所有肽段提取特征，提取的特征既包括母离子同位素峰簇相似度、母离子质量误差等单张二级谱图的打分，也包括肽段保留时间范围内的肽谱匹配打分平均值、碎片离子质量误差等，而后使用 Percolator 算法进行重排序。

3.1.4 谱库搜索方法的重排序

谱库搜索方法在构建谱库的过程中通常会引入错误的结果，如 DDA 软件报告的、公开谱库中存在的、深度学习预测的各种错误结果。不同大小的谱库中含有错误数目不同，一般来说更大的谱库所含错误数目更多，因此需要更严格的可信度评估。目前常用的重排序算法有 mProphet、PyProphet、Percolator，此外还有 DIA-NN、MaxDIA 和 DreamDIA 等软件内置的重排序算法。

mProphet 对目标库和诱饵库匹配结果计算多维特征，使用基于线性判别分析（LDA）的半监督学习方法进行二分类训练，得到各个子特征的权重并对每个谱峰组线性计算得到判别打分。PyProphet 在 mProphet 的基础上进行了重写并在半监督学习和 FDR 估计方面进行了改进，有更多可选择的机器学习模型（如 SVM、SGD 和 XGBoost 以及 LDA）。在交叉验证方面，PyProphet 使用所有数据集用于训练，替换了随机选取一半数据集分别用于训练和验证的方法。

EncyclopeDIA利用Percolator实现了重排序算法。对每个肽段结果在打分最高点计算二级谱图总体匹配打分、母离子和碎片离子准确度打分以及保留时间准确性打分共15个辅助匹配特征，通过半监督SVM模型Percolator进行重排序。Percolator 3将目标库和诱饵库结果随机分成3个子集，为每个分类器选择两份进行训练，剩余1份用于验证。通过交叉验证总共训练3个SVM分类器，并用分类器的平均值作为最终打分。Percolator 3通过下采样和交叉验证，减轻了过拟合的影响，同时提高了在大数据集上的运行速度。

DIA-NN首次通过神经网络模型实现了DIA重排序。该方法对肽段结果依据最优碎片离子分别计算基于母离子、碎片离子及其同位素离子的共洗脱曲线以及谱峰组其余谱峰的相似度等共73维打分特征。DIA-NN搭建了5层隐藏层并使用Tanh函数作为激活函数，输出层使用Softmax函数输出分类概率和交叉熵损失函数，通过输入归一化的73维特征进行学习。整个模型由73×25×20×15×10×5×1的神经网络组成，共有273.75万个神经元。DIANN使用来自目标库和诱饵库的所有鉴定结果作为训练集进行有监督学习，通过多个不同初始化参数的网络实现集成学习。

DDIA的重排序算法利用DDA扫描中的鉴定结果来划分训练集，将DDA扫描鉴定到的肽段作为参考序列，去DIA数据中进行正负例数据提取并训练分类器，用于谱库中剩余肽段的分类。对校准后的谱库结果进行靶向数据提取，并使用分类器判断得到最终肽段鉴定结果。

MaxDIA软件使用XGBoost机器学习算法进行重排序。对于每个肽段结果，MaxDIA提取了匹配结果的碎片离子相关性、保留时间误差、是否存在母离子同位素峰簇和基于碎片离子信息计算的打分、碎片离子质量误差、是否存在碎片离子同位素峰簇等60维特征，输入到XGBoost模型中进行训练。为了消除模型过度拟合的风险，作者使用5折交叉验证来训练XGBoost模型。

DreamDIA软件使用LSTM模型将肽段在DIA数据中各类型的色谱曲线转为16维高维特征，并结合保留时间差值、谱库和实际碎片离子强度相关性以及肽段的长度、电荷和质荷比等启发式特征构建非线性判别模型。在半监督学习过程中，DreamDIA采用了PU-Learning的思想[94]，选择所有诱饵库结果作为负例，根据目标库结果的打分阈值筛选正例，使用XGBoost机器学习模型进行一次训练。

DIA重排序在特征选择和训练方法上仍存在进一步优化空间，并且算法本身的可信度也需要进行有效评估。目前的DIA重排序算法都是基于传统启发式特征或经过表示学习得到的特征，使用线性或非线性机器学习模型进行训练和重打分。未来使用深度学习直接基于DIA原始数据进行建模是重排序的重要优化方向。此外，诱饵库构建、正负例选择和模型训练的差异可能会对后续分析产生影响，比如模型过拟合。因此，重排序算法本身的可信度仍然需要进一步的评估。目前还未有系统评估DIA数据重排序算法可信度的方法。

3.2　假发现率估计

肽段匹配结果重排序后，需要划定一个阈值，将满足阈值条件的结果集合报告给用户，并对该集合的可信度进行量化评估。目前广泛使用的可信度评价指标是FDR，本节先介绍FDR的定义，然后总结常见的FDR的计算过程，最后对FDR在DIA方法中现存的问题进行讨论。

FDR在蛋白质组学中表示为随机匹配结果占所有匹配结果比例的期望，即$\mathrm{FDR}(x)=\mathrm{E}[N_{\mathrm{v}}(x)/N_{\mathrm{r}}(x)]$（$x$为打分阈值，$N_{\mathrm{v}}(x)$为打分大于等于$x$的错误匹配结果数目，$N_{\mathrm{r}}(x)$为打分大于等于$x$的匹配结果数目）。

由于无法确认肽谱匹配集合中的错误匹配结果，不能直接计算得到错误匹配结果数目$N_{\mathrm{v}}(x)$。Gygi等[46]假设来自诱饵库的肽谱匹配数目和来自目标库的错误匹配数目是近似相等的，通过构建合理的诱饵库来估计$N_{\mathrm{v}}(x)$。为了估计FDR，利用诱饵库匹配结果估计目标库的错误的打分分布或数目，分别提出了基于诱饵库匹配结果打分分布估计FDR的方法和基于诱饵库匹配结果数目估计FDR的方法。mProphet、PyProphet等软件通过诱饵库匹配结果打分分布估计FDR，并使用谱库中包含的在特定样本中无法检测到肽段的比例π_0[62]来控制FDR的变化，π_0值较大表明需要对数据进行更严格的质量控制。mProphet、PyProphet使用对应诱饵库结果的打分分布拟合错误结果的打分分布，然后基于该分布对目标库结果计算P值，并通过BH算法估计FDR。mProphet和PyProphet计算q值时在原有BH算法基础上乘以系数π_0，来控制不同错误率数据集的FDR大小，具有较大π_0的数据会相应得到更严格的FDR控制。DIA-NN等软件使用

诱饵库匹配结果数目估计FDR，即使用对应诱饵库匹配的结果数目来估计错误匹配的结果数目$N_v(x)$，此时FDR的计算公式就变为$FDR(x) \approx N_d(x)/N_r(x)$（$x$为打分阈值，$N_d(x)$为来自诱饵库的打分大于等于$x$的匹配结果数目，$N_r(x)$为来自目标库的打分大于等于$x$的匹配结果数目）。

FDR作为DDA方法常用的可信度评估指标，在DIA方法中的有效性仍缺少系统评估。FDR的准确性与诱饵库的构建方式相关，目前DIA构建诱饵库的方式和DDA类似，根据目标库的肽段序列进行修改。常用的诱饵库构建方式有序列随机打乱（shuffle）、序列反转（reverse）、序列伪反转（pseudo-reverse）、质量偏移（shift）、德·布鲁因图构建（de Bruijn）等[28, 33, 95]。由于DIA方法与DDA方法在数据和解析方法上存在的差异，相同的目标-诱饵库集合经过肽段匹配可能得到不同的FDR，DDA数据的诱饵库构建结论难以直接用于DIA数据中。此外，不同DIA软件所用的诱饵库构建方法并不一致，导致各软件实际的诱饵库结果打分分布搜索空间的差异，一定程度上增加了软件之间的不可比性。综上，针对DIA数据分析，还没有FDR估计准确性的相关研究，且仍然需要探索公认合理的诱饵库构建方法。

4　总结与展望

DIA作为近年来新兴的一种数据采集技术，由于其高通量、高灵敏度、高重现性的特点，被广泛用于蛋白质组学的大规模分析[8, 96-97]，在磷酸化蛋白质组学[98]和糖蛋白质组学[74, 99]等领域也有着广泛的应用。目前分析DIA数据的主要挑战是实现对含有多个母离子共碎裂信息的混合二级谱图进行有效可靠的解析。针对这一挑战，研究人员提出了多种优化的DIA采集方法和高效的数据分析策略。近年来发展的DIA采集方法在增加蛋白质覆盖深度的同时降低了谱图复杂度。基于不同搜索策略的DIA数据解析方法对数据实现了深度解析。对于数据解析得到的肽谱匹配，对其进行重排序和假发现率估计，最终获取高可信的肽谱匹配集合。DIA数据定量分析通过重构碎片离子色谱曲线计算肽段强度（图5），较DDA定量具有更好的定量精度。此外，DIA数据具有定性定量一致性的特点，通过对碎片离子色谱曲线进行肽段打分，能够对数据之间对齐（match between runs）的结果进行可信度评估。DIA定量具有深度覆盖、可重现性和定量精确性等优点，使用DIA对大规模生物样品进行定量是实现蛋白质深度覆盖的发展趋势。

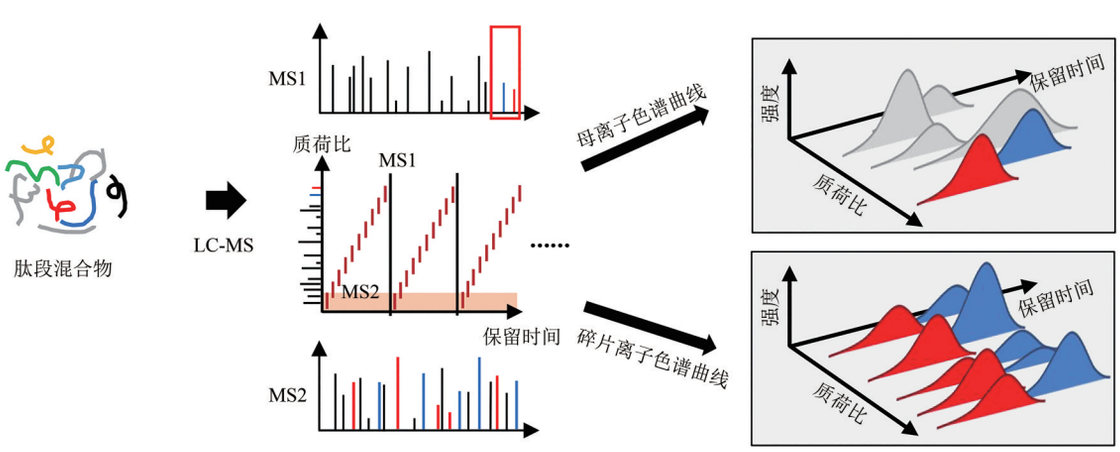


Fig. 5　The XIC of fragment ion was reconstructed from DIA data

图5　DIA数据重构碎片离子色谱曲线

DIA方法虽然在近年来获得了快速的发展，且在深度覆盖方面较DDA有更好的表现，但是在深度解析、精准鉴定和精准定量3个方面仍存在不足和进步空间。在深度解析方面，DIA数据理论上采集了样品内所有肽段的碎片离子信息，能够通过谱图解析实现蛋白质完整肽段的鉴定。但是，对二级谱图实现深度解析需要考虑意外酶切情况和对低丰度肽段、多种类型修饰和意外修饰等非常规肽段的鉴定，同时也要考虑DIA数据存在的肽段共碎裂造成的离子干扰等情况。如果直接通过扩大搜索空

间搜索潜在候选肽段不仅会增加搜索的时间，也可能会影响鉴定结果可靠性。因此，如何在考虑包含更多潜在肽段搜索空间的同时，设计更有效的谱图解析算法提高非常规肽段的鉴定数目，同时能保证搜索时间在合理范围内，是DIA数据实现深度解析目前面临的一个主要挑战。

在精准鉴定方面，虽然利用TDA方法计算FDR的策略为DIA数据分析结果的质量控制提供了可能，但是该策略在DIA数据的有效性目前还没有定论。首先，不同的诱饵库构建方法会对结果造成影响，如何使用和设计能更好评估结果假发现率的诱饵库仍需进一步系统评估。其次，肽段在DIA数据中是以保留时间上连续的二级谱图序列形式呈现的，在实际匹配过程中可能会出现不同肽段匹配到大量相同谱峰的竞争情况，而传统FDR评估方法很少考虑到不同肽段鉴定结果之间的竞争情况，如何针对此类情况设计合适的质量控制方法是未来需要研究的课题之一。最后，由于DIA数据采集了不同梯度的肽段，对所有结果统一进行质量控制可能是不合适的。如低丰度肽段在数据中可能受到离子干扰、离子抑制和噪声干扰等情况，可能会造成匹配打分不高而被视作错误结果排除。如何针对不同类型的肽段鉴定结果和其匹配情况设计分层次的质量控制是未来需要研究的课题之一。

在精准定量方面，DIA数据分析在定量缺失值和定量精度方面取得了较好的表现，但是仍有进一步提升空间。一方面，在通过数据之间对齐降低定量缺失值时，可以考虑母离子和碎片离子色谱曲线来优化肽段信号的匹配，同时利用该信息进一步评估定量结果的正确性。另一方面，可以利用母离子和碎片离子色谱曲线的全面信息进一步提高定量精度，更好地结合母离子和碎片离子各自的定量优势，并进一步利用该信息去除离子干扰情况，实现精准定量。

随着质谱采集的优化和数据分析的发展，DIA采集技术在进一步解决上述介绍的不足后，能够为蛋白质组学的高通量、全覆盖分析提供进一步的支持，特别是在大队列数据分析中均能获取完整蛋白质图谱并解释其潜在生命规律，推动蛋白质组学领域的发展。利用DIA采集技术，可以建立包含所有肽段和蛋白质信息的数字化标本库，实现数以千计、万计的样品的蛋白质组学深度解析以及横向比较，并结合先进的人工智能技术，进一步进行深度数据挖掘，发现更有效的疾病标志物，探索更深层次的分子细胞作用机制，为生命科学及人类健康研究做出重大贡献。

参 考 文 献

[1] Steen H, Mann M. The ABC’s (and XYZ’s) of peptide sequencing. Nat Rev Mol Cell Biol, 2004, **5**(9): 699-711

[2] Cottrell J S. Protein identification using MS/MS data. J Proteomics, 2011, **74**(10): 1842-1851

[3] Eckert M A, Coscia F, Chryplewicz A, *et al*. Proteomics reveals NNMT as a master metabolic regulator of cancer-associated fibroblasts. Nature, 2019, **569**(7758): 723-728

[4] Tian W, Zhang N, Jin R, *et al*. Immune suppression in the early stage of COVID-19 disease. Nat Commun, 2020, **11**(1): 5859

[5] Su M, Zhang Z, Zhou L, *et al*. Proteomics, personalized medicine and cancer. Cancers (Basel), 2021, **13**(11): 2512

[6] Bassani-Sternberg M, Braunlein E, Klar R, *et al*. Direct identification of clinically relevant neoepitopes presented on native human melanoma tissue by mass spectrometry. Nat Commun, 2016, **7**: 13404

[7] Xu J Y, Zhang C, Wang X, *et al*. Integrative proteomic characterization of human lung adenocarcinoma. Cell, 2020, **182**(1): 245-261.e17

[8] Dyring-Andersen B, Lovendorf M B, Coscia F, *et al*. Spatially and cell-type resolved quantitative proteomic atlas of healthy human skin. Nat Commun, 2020, **11**(1): 5587

[9] Jiang L, Wang M, Lin S, *et al*. A quantitative proteome map of the human body. Cell, 2020, **183**(1): 269-283.e19

[10] Nie X, Qian L, Sun R, *et al*. Multi-organ proteomic landscape of COVID-19 autopsies. Cell, 2021, **184**(3): 775-791.e14

[11] Bai B, Wang X, Li Y, *et al*. Deep multilayer brain proteomics identifies molecular networks in Alzheimer’s disease progression. Neuron, 2020, **106**(4): 700

[12] Rosenberger G, Koh C C, Guo T, *et al*. A repository of assays to quantify 10,000 human proteins by SWATH-MS. Sci Data, 2014, **1**: 140031

[13] Adhikari S, Nice E C, Deutsch E W, *et al*. A high-stringency blueprint of the human proteome. Nat Commun, 2020, **11**(1): 5301

[14] Venable J D, Dong M Q, Wohlschlegel J, *et al*. Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra. Nat Methods, 2004, **1**(1): 39-45

[15] Purvine S, Eppel J T, Yi E C, *et al*. Shotgun collision-induced dissociation of peptides using a time of flight mass analyzer. Proteomics, 2003, **3**(6): 847-850

[16] Plumb R S, Johnson K A, Rainville P, *et al*. UPLC/MS(E); a new approach for generating molecular fragment information for biomarker structure elucidation. Rapid Commun Mass Spectrom, 2006, **20**(13): 1989-1994

[17] Geiger T, Cox J, Mann M. Proteomics on an Orbitrap benchtop mass spectrometer using all-ion fragmentation. Mol Cell Proteomics, 2010, **9**(10): 2252-2261

[18] Panchaud A, Scherl A, Shaffer S A, *et al*. Precursor acquisition

independent from ion count: how to dive deeper into the proteomics ocean. Anal Chem, 2009, **81**(15): 6481-6488

[19] 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**(6): O111.016717

[20] Martin L B, Sherwood R W, Nicklay J J, *et al*. Application of wide selected-ion monitoring data-independent acquisition to identify tomato fruit proteins regulated by the CUTIN DEFICIENT2 transcription factor. Proteomics, 2016, **16**(15-16): 2081-2094

[21] Meier F, Geyer P E, Virreira Winter S, *et al*. BoxCar acquisition method enables single-shot proteomics at a depth of 10, 000 proteins in 100 minutes. Nat Methods, 2018, **15**(6): 440-448

[22] Egertson J D, Kuehn A, Merrihew G E, *et al*. Multiplexed MS/MS for improved data-independent acquisition. Nat Methods, 2013, **10**(8): 744-746

[23] Zhang Y, Bilbao A, Bruderer T, *et al*. The use of cariable Q1 isolation windows improves selectivity in LC-SWATH-MS acquisition. J Proteome Res, 2015, **14**(10): 4359-4371

[24] Li W, Chi H, Salovska B, *et al*. Assessing the relationship between mass window width and retention time scheduling on protein coverage for data-independent acquisition. J Am Soc Mass Spectrom, 2019, **30**(8): 1396-1405

[25] Meier F, Brunner A D, Frank M, *et al*. diaPASEF: parallel accumulation-serial fragmentation combined with data-independent acquisition. Nat Methods, 2020, **17**(12): 1229-1236

[26] Messner C B, Demichev V, Bloomfield N, *et al*. Ultra-fast proteomics with Scanning SWATH. Nat Biotechnol, 2021, **39**(7): 846-854

[27] Reiter L, Rinner O, Picotti P, *et al*. mProphet: automated data processing and statistical validation for large-scale SRM experiments. Nat Methods, 2011, **8**(5): 430-435

[28] Rost H L, Rosenberger G, Navarro P, *et al*. OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. Nat Biotechnol, 2014, **32**(3): 219-223

[29] Bruderer R, Bernhardt O M, Gandhi T, *et al*. Extending the limits of quantitative proteome profiling with data-independent acquisition and application to acetaminophen-treated three-dimensional liver microtissues. Mol Cell Proteomics, 2015, **14**(5): 1400-1410

[30] Teleman J, Rost H L, Rosenberger G, *et al*. DIANA--algorithmic improvements for analysis of data-independent acquisition MS data. Bioinformatics, 2015, **31**(4): 555-562

[31] Keller A, Bader S L, Shteynberg D, *et al*. Automated validation of results and removal of fragment ion interferences in targeted analysis of data-independent acquisition mass spectrometry (MS) using SWATHProphet. Mol Cell Proteomics, 2015, **14**(5): 1411-1418

[32] Searle B C, Pino L K, Egertson J D, *et al*. Chromatogram libraries improve peptide detection and quantification by data independent acquisition mass spectrometry. Nat Commun, 2018, **9**(1): 5128

[33] Demichev V, Messner C B, Vernardis S I, *et al*. DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput. Nat Methods, 2020, **17**(1): 41-44

[34] Guan S, Taylor P P, Han Z, *et al*. Data dependent-independent acquisition (DDIA) proteomics. J Proteome Res, 2020, **19**(8): 3230-3237

[35] Sinitcyn P, Hamzeiy H, Salinas Soto F, *et al*. MaxDIA enables library-based and library-free data-independent acquisition proteomics. Nat Biotechnol, 2021, **39**(12):1563-1573

[36] Gao M, Yang W, Li C, *et al*. Deep representation features from DreamDIA(XMBD) improve the analysis of data-independent acquisition proteomics. Commun Biol, 2021, **4**(1): 1190

[37] Lu Y Y, Bilmes J, Rodriguez-Mias R A, *et al*. DIAmeter: matching peptides to data-independent acquisition mass spectrometry data. Bioinformatics, 2021, **37**(Suppl_1): i434-i442

[38] Weisbrod C R, Eng J K, Hoopmann M R, *et al*. Accurate peptide fragment mass analysis: multiplexed peptide identification and quantification. J Proteome Res, 2012, **11**(3): 1621-1632

[39] Ting Y S, Egertson J D, Bollinger J G, *et al*. PECAN: library-free peptide detection for data-independent acquisition tandem mass spectrometry data. Nat Methods, 2017, **14**(9): 903-908

[40] Bern M, Finney G, Hoopmann M R, *et al*. Deconvolution of mixture spectra from ion-trap data-independent-acquisition tandem mass spectrometry. Anal Chem, 2010, **82**(3): 833-841

[41] Tsou C C, Avtonomov D, Larsen B, *et al*. DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics. Nat Methods, 2015, **12**(3): 258-264

[42] Li Y, Zhong C Q, Xu X, *et al*. Group-DIA: analyzing multiple data-independent acquisition mass spectrometry data files. Nat Methods, 2015, **12**(12): 1105-1106

[43] Peckner R, Myers S A, Jacome A S V, *et al*. Specter: linear deconvolution for targeted analysis of data-independent acquisition mass spectrometry proteomics. Nat Methods, 2018, **15**(5): 371-378

[44] Tada I, Chaleckis R, Tsugawa H, *et al*. Correlation-based deconvolution (CorrDec) to generate high-quality MS2 spectra from data-independent acquisition in multisample studies. Anal Chem, 2020, **92**(16): 11310-11317

[45] Tran N H, Qiao R, Xin L, *et al*. Deep learning enables *de novo* peptide sequencing from data-independent-acquisition mass spectrometry. Nat Methods, 2019, **16**(1): 63-66

[46] Elias J E, Gygi S P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. Nat Methods, 2007, **4**(3): 207-214

[47] Kall L, Storey J D, Maccoss M J, *et al*. Posterior error probabilities and false discovery rates: two sides of the same coin. J Proteome Res, 2008, **7**(1): 40-44

[48] Amodei D, Egertson J, Maclean B X, *et al*. Improving precursor selectivity in data-independent acquisition using overlapping windows. J Am Soc Mass Spectrom, 2019, **30**(4): 669-684

[49] Heaven M R, Funk A J, Cobbs A L, *et al*. Systematic evaluation of data-independent acquisition for sensitive and reproducible proteomics-a prototype design for a single injection assay. J Mass

Spectrom, 2016, **51**(1): 1-11

[50] Cai X, Ge W, Yi X, *et al*. PulseDIA: data-independent acquisition mass spectrometry using multi-injection pulsed gas-phase fractionation. J Proteome Res, 2021, **20**(1): 279-288

[51] Kiyonami R, Patel B, Senko M, *et al*. Large Scale Targeted Protein Quantification Using WiSIM-DIA Workflow on a Orbitrap Fusion Tribrid Mass Spectrometer. Waltham: Thermo Fisher, 2014

[52] Xuan Y, Bateman N W, Gallien S, *et al*. Standardization and harmonization of distributed multi-center proteotype analysis supporting precision medicine studies. Nat Commun, 2020, **11**(1): 5248

[53] Mehta D, Scandola S, Uhrig R G J B. Library-free BoxCarDIA solves the missing value problem in label-free quantitative proteomics. bioRxiv, 2021. doi: 10.1101/2020.11.07.372276

[54] Salovska B, Li W, Di Y, *et al*. BoxCarmax: a high-selectivity data-independent acquisition mass spectrometry method for the analysis of protein turnover and complex samples. Anal Chem, 2021, **93**(6): 3103-3111

[55] Meier F, Park M A, Mann M. Trapped ion mobility spectrometry and parallel accumulation-serial fragmentation in proteomics. Mol Cell Proteomics, 2021, **20**: 100138

[56] Yates J R, 3rd, Morgan S F, Gatlin C L, *et al*. Method to compare collision-induced dissociation spectra of peptides: potential for library searching and subtractive analysis. Anal Chem, 1998, **70**(17): 3557-3565

[57] Lam H, Deutsch E W, Eddes J S, *et al*. Development and validation of a spectral library searching method for peptide identification from MS/MS. Proteomics, 2007, **7**(5): 655-667

[58] Lam H, Deutsch E W, Eddes J S, *et al*. Building consensus spectral libraries for peptide identification in proteomics. Nat Methods, 2008, **5**(10): 873-875

[59] Lam H, Deutsch E W, Aebersold R. Artificial decoy spectral libraries for false discovery rate estimation in spectral library searching in proteomics. J Proteome Res, 2010, **9**(1): 605-610

[60] Bruderer R, Bernhardt O M, Gandhi T, *et al*. Optimization of experimental parameters in data-independent mass spectrometry significantly increases depth and reproducibility of results. Mol Cell Proteomics, 2017, **16**(12): 2296-2309

[61] Parker S J, Venkatraman V, Van Eyk J E. Effect of peptide assay library size and composition in targeted data-independent acquisition-MS analyses. Proteomics, 2016, **16**(15-16): 2221-2237

[62] Rosenberger G, Bludau I, Schmitt U, *et al*. Statistical control of peptide and protein error rates in large-scale targeted data-independent acquisition analyses. Nat Methods, 2017, **14**(9): 921-927

[63] Xu L L, Young A, Zhou A, *et al*. Machine learning in mass spectrometric analysis of DIA data. Proteomics, 2020, **20**(21-22): e1900352

[64] Wen B, Zeng W F, Liao Y, *et al*. Deep learning in proteomics. Proteomics, 2020, **20**(21-22): e1900335

[65] Mann M, Kumar C, Zeng W F, *et al*. Artificial intelligence for proteomics and biomarker discovery. Cell Syst, 2021, **12**(8): 759-770

[66] Zhou X X, Zeng W F, Chi H, *et al*. pDeep: predicting MS/MS spectra of peptides with deep learning. Anal Chem, 2017, **89**(23): 12690-12697

[67] Gessulat S, Schmidt T, Zolg D P, *et al*. Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. Nat Methods, 2019, **16**(6): 509-518

[68] Yang Y, Liu X, Shen C, *et al*. In silico spectral libraries by deep learning facilitate data-independent acquisition proteomics. Nat Commun, 2020, **11**(1): 146

[69] Maclean B, Tomazela D M, Shulman N, *et al*. Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. Bioinformatics, 2010, **26**(7): 966-968

[70] Midha M K, Campbell D S, Kapil C, *et al*. DIALib-QC an assessment tool for spectral libraries in data-independent acquisition proteomics. Nat Commun, 2020, **11**(1): 5251

[71] Zhu T, Zhu Y, Xuan Y, *et al*. DPHL: a DIA Pan-human protein mass spectrometry library for robust biomarker discovery. Genomics Proteomics Bioinformatics, 2020, **18**(2): 104-119

[72] Wang D, Gan G, Chen X, *et al*. QuantPipe: a user-friendly pipeline software tool for DIA data analysis based on the OpenSWATH-PyProphet-TRIC workflow. J Proteome Res, 2021, **20**(1): 1096-1102

[73] Bichmann L, Gupta S, Rosenberger G, *et al*. DIAproteomics: a multifunctional data analysis pipeline for data-independent acquisition proteomics and peptidomics. J Proteome Res, 2021, **20**(7):3758-3766

[74] Yang Y, Yan G, Kong S, *et al*. GproDIA enables data-independent acquisition glycoproteomics with comprehensive statistical control. Nat Commun, 2021, **12**(1): 6073

[75] Pino L K, Searle B C, Bollinger J G, *et al*. The skyline ecosystem: informatics for quantitative mass spectrometry proteomics. Mass Spectrom Rev, 2020, **39**(3): 229-244

[76] Finney G L, Blackler A R, Hoopmann M R, *et al*. Label-free comparative analysis of proteomics mixtures using chromatographic alignment of high-resolution μLC-MS data. 2008, **80**(4): 961-971

[77] Demichev V, Yu F, Teo G C, *et al*. High sensitivity dia-PASEF proteomics with DIA-NN and FragPipe. bioRxiv, 2021. doi: 10.1101/2021.03.08.434385

[78] Tiwary S, Levy R, Gutenbrunner P, *et al*. High-quality MS/MS spectrum prediction for data-dependent and data-independent acquisition data analysis. Nat Methods, 2019, **16**(6): 519-525

[79] Wang J, Tucholska M, Knight J D, *et al*. MSPLIT-DIA: sensitive peptide identification for data-independent acquisition. Nat Methods, 2015, **12**(12): 1106-1108

[80] Zhang X, Li Y, Shao W, *et al*. Understanding the improved sensitivity of spectral library searching over sequence database searching in proteomics data analysis. Proteomics, 2011, **11**(6): 1075-1085

[81] Fernandez-Costa C, Martinez-Bartolome S, Mcclatchy D B, *et al*.

Impact of the identification strategy on the reproducibility of the DDA and DIA results. J Proteome Res, 2020, **19**(8): 3153-3161

[82] Ye D, Fu Y, Sun R X, *et al*. Open MS/MS spectral library search to identify unanticipated post-translational modifications and increase spectral identification rate. Bioinformatics, 2010, **26**(12): i399-406

[83] Anderson D C, Li W, Payan D G, *et al*. A new algorithm for the evaluation of shotgun peptide sequencing in proteomics: support vector machine classification of peptide MS/MS spectra and SEQUEST scores. J Proteome Res, 2003, **2**(2): 137-146

[84] Venable J D, Xu T, Cociorva D, *et al*. Cross-correlation algorithm for calculation of peptide molecular weight from tandem mass spectra. Anal Chem, 2006, **78**(6): 1921-1929

[85] Zhang N, Li X J, Ye M, *et al*. ProbIDtree: an automated software program capable of identifying multiple peptides from a single collision-induced dissociation spectrum collected by a tandem mass spectrometer. Proteomics, 2005, **5**(16): 4096-4106

[86] Sulimov P, Kertesz-Farkas A. Tailor: a nonparametric and rapid score calibration method for database search-based peptide identification in shotgun proteomics. J Proteome Res, 2020, **19**(4): 1481-1490

[87] Ting Y S, Egertson J D, Payne S H, *et al*. Peptide-centric proteome analysis: an alternative strategy for the analysis of tandem mass spectrometry data. Mol Cell Proteomics, 2015, **14**(9): 2301-2307

[88] Craig R, Cortens J P, Beavis R C. Open source system for analyzing, validating, and storing protein identification data. J Proteome Res, 2004, **3**(6): 1234-1242

[89] Eng J K, Jahan T A, Hoopmann M R. Comet: an open-source MS/MS sequence database search tool. Proteomics, 2013, **13**(1): 22-24

[90] Kim S, Mischerikow N, Bandeira N, *et al*. The generating function of CID, ETD, and CID/ETD pairs of tandem mass spectra: applications to database search. Mol Cell Proteomics, 2010, **9**(12): 2840-2852

[91] Sadygov R G, Maroto F M, Huhmer A F. ChromAlign: a two-step algorithmic procedure for time alignment of three-dimensional LC-MS chromatographic surfaces. Anal Chem, 2006, **78**(24): 8207-8217

[92] Tsugawa H, Ikeda K, Takahashi M, *et al*. A lipidome atlas in MS-DIAL 4. Nat Biotechnol, 2020, **38**(10): 1159-1163

[93] Kall L, Canterbury J D, Weston J, *et al*. Semi-supervised learning for peptide identification from shotgun proteomics datasets. Nat Methods, 2007, **4**(11): 923-925

[94] Elkan C P, Noto K. Learning classifiers from only positive and unlabeled data. Proceedings of the KDD, 2008, **213**: 220

[95] Moosa J M, Guan S, Moran M F, *et al*. Repeat-preserving decoy database for false discovery rate estimation in peptide identification. J Proteome Res, 2020, **19**(3): 1029-1036

[96] Guo T, Kouvonen P, Koh C C, *et al*. Rapid mass spectrometric conversion of tissue biopsy samples into permanent quantitative digital proteome maps. Nat Med, 2015, **21**(4): 407-413

[97] Lou R, Liu W, Li R, *et al*. DeepPhospho accelerates DIA phosphoproteome profiling through in silico library generation. Nat Commun, 2021, **12**(1): 6685

[98] Bekker-Jensen D B, Bernhardt O M, Hogrebe A, *et al*. Rapid and site-specific deep phosphoproteome profiling by data-independent acquisition without the need for spectral libraries. Nat Commun, 2020, **11**(1): 787

[99] Ye Z, Mao Y, Clausen H, *et al*. Glyco-DIA: a method for quantitative O-glycoproteomics with in silico-boosted glycopeptide libraries. Nat Methods, 2019, **16**(9): 902-910

Progress in Data Analysis Methods for Proteome Mass Spectrometry Based on Data-independent Acquisition*

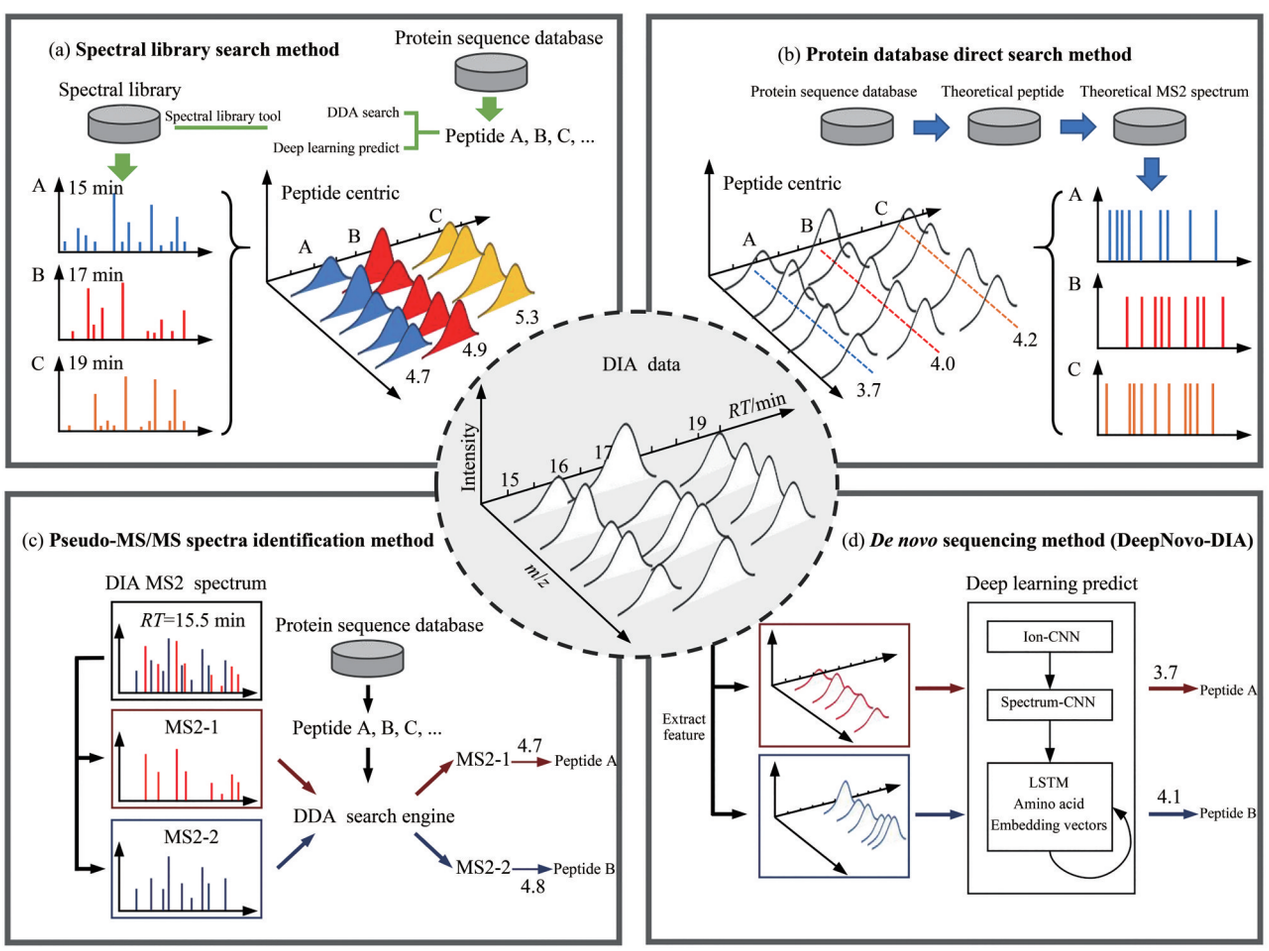
HOU Xin-Hang[1)], ZHOU Pi-Yu[1)], GONG Peng-Yun[2)], FU Jia-Le[3)], LIU Chao[2)**], WANG Hai-Peng[1)**]

([1)]*School of Computer Science and Technology, Shandong University of Technology, Zibo* 255000, China;

[2)]*School of Engineering Medicine & School of Biological Sicence and Medical Engineering, Beihang University, Beijing* 100191, China;

[3)]*School of Life Sciences, Tsinghua University, Beijing* 100084, China)

Graphical abstract



The workflow of DIA data analysis method

* This work was supported by grants from Support Program for Outstanding Youth Innovation Teams in Higher Educational Institutions of Shandong Province (2019KJN048) and The National Natural Science Foundation of China (31500669).

** Corresponding author.

WANG Hai-Peng. Tel：86-533-2783479，E-mail：hpwang@sdut.edu.cn

LIU Chao. Tel：86-10-82316427，E-mail：liuchaobuaa@buaa.edu.cn

Received：November 16，2021　Accepted：March 21，2022

Abstract Data independent acquisition (DIA) is a rapidly developing proteomics technique in recent years, which can theoretically achieve deep coverage of protein samples by collecting tandem mass spectra through unbiased co-fragmentation of all precursors in the isolation window. It has the advantages of high throughput, high reproducibility and high sensitivity. Current DIA data acquisition methods mainly include full-window fragmentation method, isolation window sequential fragmentation method and four-dimensional DIA data acquisition method (4D-DIA). The most commonly used data acquisition methods are SWATH or variable window SWATH and DIA-PASEF methods. The tandem mass spectra collected by the full-window fragmentation method contains precursor ions in the full *m*/*z* range, and the spectra analysis is complex. The isolation window sequential fragmentation method reduces the number of precursor ions in tandem mass spectra and the size of the isolation window through a variety of acquisition strategies, effectively reducing the complexity of spectra interpretation. With the development of mass spectrometry instruments, the size of isolation window of the tandem mass spectra acquired by DIA may be close to the size of DDA, enabling the integration of DIA and DDA processes. The 4D-DIA method obtains the corresponding relationship between precursor and fragment ions through additional data dimensions, which improve the selectivity of precursor and greatly reduce the complexity of spectral analysis. The 4D-DIA method is also an important advance for future DIA data collection. According to the characteristics of DIA data, relevant data analysis methods were designed, which mainly included spectral library search method, protein database direct search method, pseudo-MS/MS spectra identification method and *de novo* sequencing method, as showed in the figure above. The spectral library search method uses the spectral library information for data extraction, which has high peptide identification sensitivity, but have certain requirements on the quality and number of spectral libraries; the protein database direct search method does not require preprocessing of tandem mass spectra and construction of spectral libraries, and directly matches the theoretical tandem mass spectrum of peptide with experimental tandem mass spectrum, but the time complexity is high; pseudo-MS/MS spectra identification method uses the spectra splitting algorithm to split the tandem mass spectrum to obtain multiple pseudo-MS/MS spectra containing single peptide fragment ions, then combined with traditional DDA software to search pseudo-MS/MS spectra; *de novo* sequencing method directly models the pseudo-MS/MS spectrum through deep learning to predict peptides, has the advantage of identifying sequences of new species, but it is difficult to guarantee the number and reliability of the identification results. The reliability evaluation of the peptide-spectrum matches mainly includes re-ranking by machine learning and false discovery rate estimation of the reported results. Although the DIA method has achieved rapid development in recent years, and has better performance than DDA in terms of depth coverage, there are still shortcomings and improvement in 3 aspects: in-depth analysis, accurate identification and accurate quantification. With the optimization of mass spectrometry acquisition and the development of data analysis, DIA acquisition technology can provide further support for high throughput, full-coverage analysis of proteomics, especially in large cohort data analysis, after further solving the above-mentioned shortcomings. All of them can obtain complete protein maps and explain their underlying life laws, promoting the development of the field of proteomics. In this paper, the DIA data collection method, data analysis method, software and identification result reliability assessment method are sorted and reviewed, and the future development direction is prospected.

Key words proteomics, data independent acquisition, mass spectrometry, targeted data extraction
DOI：10.16476/j.pibb.2021.0345