PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2024, 51(12): 3151~3162
www.pibb.ac.cn

蛋白质组学数据揭示可变剪接蛋白质变体*

吴怡颖　张　炜**　孔德志**
（河北医科大学中西医结合研究所，石家庄 050017）

摘要　可变剪接使同一个基因产生多种不同的转录本和蛋白质，增加蛋白质多样性和功能复杂性。转录组学和蛋白质组学是鉴定可变剪接的两种主要途径，分别通过分析RNA测序数据和蛋白质测序数据来鉴定可变剪接事件和相应蛋白质产物。尽管RNA水平的可变剪接研究已经取得了一定的进展，但是对于剪接蛋白亚型的检测和区分仍然不足。本文综述了近年来对可变剪接及其生物功能，可变剪接在不同水平检测的研究进展，详细介绍利用“自下而上”的蛋白质组学数据鉴定可变剪接蛋白质变体的两种主要方法，序列数据库构建和蛋白质鉴定算法开发。鉴定不同的可变剪接蛋白质有利于认识更全面的蛋白质功能，对发现相关生物标志物和关键药物靶点具有重要意义。

关键词　可变剪接，质谱数据分析，蛋白质鉴定算法，蛋白质序列数据库
中图分类号　Q51，R917　　**DOI**：10.16476/j.pibb.2024.0109

随着转录组学和蛋白质组学等技术的发展和应用，我们可以更深入地研究和理解生命的复杂性和多样性，这些高通量和系统性的方法在疾病机理研究、药物靶点发现和治疗策略开发等方面发挥着关键作用。同时，组学技术和组学数据分析在向着更加精细化的方向发展，例如，单细胞测序技术（single-cell sequencing，SCS）[1]的发展使基因组学从大量组织分析转向对单个细胞的详细和全面的研究，RNA测序技术不仅用于检测基因转录，还用于研究剪接变异和RNA编辑事件，蛋白质组学也致力于获得更精细准确的人类蛋白质图像。

可变剪接（又称选择性剪接，alternative splicing，AS）是一种普遍且重要的生物过程[2]，它可以产生不同mRNA并影响对应蛋白质产物，在生物体的各种生理和病理过程中起着重要的作用。近年来，可变剪接已成为生物学领域的热点问题，在临床医学、农业科学[3]等多个领域也引起了广泛关注。2017年，冷泉港实验室举办了以“发现mRNA剪接四十年的研究进展”为主题的会议，进一步突显了可变剪接在科学研究中的重要地位。从鉴定典型的蛋白质发展到鉴定各类蛋白质变体[4]（proteoform，描述一个基因由于遗传变异、可变剪接以及翻译后修饰等各种基因或蛋白质的加工事件所产生的所有不同分子形式的蛋白质[5]），也成为了蛋白质组学的趋势之一。尽管利用高通量组学数据研究可变剪接催生了许多新方法和平台，但总的来说依然存在许多挑战。

本文将介绍可变剪接的基本概念及其在生物体中的重要作用，检测可变剪接事件的两种水平，重点探讨利用蛋白质组学数据来鉴定和分析可变剪接事件的基本方法，如何利用质谱数据鉴定出可变剪接蛋白。通过对蛋白质可变剪接的深入研究，我们将更好地理解生命的复杂性，并为疾病的诊断和治疗提供新的思路。

1　可变剪接及其生物功能

1.1　可变剪接

剪接是指从前体mRNA中去除内含子并将剩余的外显子相互连接产生成熟mRNA的过程，不同外显子组合产生多种成熟mRNA的现象称为可

* 国家自然科学基金（82174004）和河北省自然科学基金（H2022206211，H2022206387）资助项目。
** 通讯联系人。
孔德志 Tel：0311-86266722，E-mail：kongdezhi@hebmu.edu.cn
张炜 Tel：0311-86266222，E-mail：weizhang@hebmu.edu.cn
收稿日期：2024-03-18，接受日期：2024-07-01

变剪接[6]。剪接过程由剪接体负责[7]，它是一种由5个小核核糖核蛋白（small nuclear ribonucleoprotein，snRNP）和大约100个蛋白质组成的大核糖核蛋白复合物[8-9]，剪接体经过识别、组装、剪切、连接和拆卸等一系列精确步骤完成工作[10]。外显子跳跃（成熟mRNA不包含某外显子）、内含子保留（内含子保留在成熟mRNA分子中）、可变的5'和3'剪接位点（外显子一端有不止一个剪接位点）、可变的5'非翻译区和3'非翻译区（剪接出现在非翻译区）、互斥外显子（两个或多个外显子不同时出现在一个mRNA中）是可变剪接的主要类型[11]，其中外显子跳跃是哺乳动物细胞中最常见的可变剪接模式[12]。可变剪接的过程普遍存在于真核生物中，人体组织中大约95%的多外显子转录本经历了可变剪接[13]，在不同的组织或者发育的不同阶段可能会产生特定的剪切异构体，具有组织和时间特异性[14]，并且这个过程受到严格的调控，当异常剪接事件发生时会影响正常的生命活动[15]。

1.2 可变剪接的生物功能

可变剪接增加基因的表达多样性，一个基因产生多种mRNA转录本，这些不同的mRNA进一步翻译成多种蛋白质亚型，同翻译后修饰一样增加蛋白质的丰富性。不同的蛋白质亚型在结构和功能上存在差异，例如，BCL-X存在两种功能相反的剪接蛋白亚型，一种保护细胞，另一种使细胞凋亡[16]，剪接会根据环境变化动态调节两种异构体比例，失调时就可能诱发一些疾病，以及由于可变剪接产生的蛋白质截短，导致分子完整性的丧失，进而影响蛋白质正常生物学功能[17]。同时可变剪接影响蛋白质与蛋白质相互作用、调节蛋白质和其他大分子[18]和小分子[19]相互作用结构域或结合位点[20]。剪接失调是阿尔茨海默病、肌萎缩侧索硬化症、额颞叶痴呆、帕金森病和重复扩张性疾病等神经退行性疾病的重要发病机制[11]；一些研究表明，肿瘤和可变剪接有关，异常剪接被认是致癌驱动因素并伴随着肿瘤的发展过程[21]，异常剪接体也被作为治疗癌症的新靶点[22]。总之，可变剪接具有重要的生物学意义，对剪接蛋白亚型的检测和表征对于揭示生物系统的基本工作机制至关重要。

2 可变剪接的检测

可变剪接是mRNA成熟过程发生的变化，在转录水平进行检测是最直接的手段，基本策略是将RNA测序技术得到的转录本序列与参考序列进行对比鉴定差异。同时，不是所有的候选转录本都会被翻译成蛋白质，直接检测可变剪接的蛋白质产物具有重要意义。

2.1 RNA水平鉴定可变剪接

现有的RNA测序技术主要有短读长测序和长读长测序。短读长测序技术测序RNA小片段（大约几百bp）；长读长测序实现较长的读长（PacBio平台最大读取长度>8 kb）[23]，能提供更完整的转录本信息。人类K562细胞转录组学数据显示，短读长测序相比长读长测序遗漏了很大一部分可变剪接情况[24]，长读长测序技术加强了对可变剪接多样性的认识，但局限性在于测序准确率低[25]。通过转录组检测可变剪接事件的主要过程是将样本的RNA测序数据映射到带注释的参考基因组，通过对比参考基因组进行识别[26]。另外，在映射步骤前应对读段进行质量控制和计数，去除低质量的读数。

工具Hercules、CoLoRMap、FLAS、LoRMA等[27]用于长读长测序的纠错和矫正；已经开发的各种对齐映射算法如FANSe3[28]、STAR[29]和Supersplat[30]等实现检测剪接事件功能，这些算法也随着测序技术的发展而改进[31]；含有转录本丰度估计功能的软件Salmon[32]；NMD分类器[33]可以识别具有提前终止密码子的RNA序列，去除那些会发生无义介导的mRNA衰变的剪接变体，达到对转录本的筛选和质控。表1列举了一些用于转录组识别可变剪接的软件。

除了高通量测序方法，常用逆转录聚合酶链式反应（reverse transcription-polymerase chain reaction，RT-PCR）实验[34-35]在转录本水平验证可变剪接产物，通过设计特异性引物，可以对预定的剪接位点进行扩增，检测特定的剪接变体是否存在，使用原位杂交技术（*in situ* hybridization，ISH）可以检测可变剪接的表达模式[36]。

2.2 蛋白质水平鉴定可变剪接

可变剪接在蛋白质水平的鉴定依靠基于质谱的蛋白质组学方法，其主要实验策略包括两种。一种是bottom-up（自下而上），bottom-up蛋白质组学通过蛋白质酶解产生的肽段来表征蛋白质，当对蛋白质混合物而不是单一蛋白质进行自下而上的综合分析时称为鸟枪法蛋白质组学[37]。另一种top-down（自顶向下）蛋白质组学分析对象是完整的

Table 1　Software for identifying alternative splicing in transcriptome data and their features

表1　转录组数据识别可变剪接的常用软件及其特点

工具	特点	代码获取
STAR	准确率高，速度快	https://github.com/alexdobin/STAR/releases
2passtools	针对长读长测序，两次对齐	https://github.com/bartongroup/2passtools
PSI-Sigma	通过剪接百分比指数（percent splicing index，PSI）值检测剪接事件	https://github.com/wososa/PSI-Sigma
Supersplat	剪接点搜索详细	http://mocklerlab.org/tools/1/manual
rMATS	针对重复的RNA测序数据	http://rnaseq-mats.sourceforge.net/
SUPPA	分析大型数据集快速准确	https://bitbucket.org/regulatorygenomicsupf/suppa
LeafCutter	提供剪接数量性状基因座图谱	https://github.com/davidaknowles/leafcutter
DiffSplice	采用基于图论的方法来分析可变剪接	http://www.netlab.uky.edu/p/bioinfo/DiffSplice
MARVEL	用于单细胞RNA测序数据	https://cloud.rproject.org/web/packages/MARVEL/index.html

蛋白质，它保留了蛋白质的完整性，可以克服bottom-up方法存在的肽段丢失与匹配歧义的情况，更容易观察到剪接、截短等相关信息[38-39]，被认为是表征翻译后修饰等蛋白质变体的有力手段[40-41]，并有学者建议检测剪接蛋白亚型使用top-down方法[42]。但通量低、灵敏度低、技术受限等问题导致top-down发展时间晚，实际应用面临较大挑战[38]。

基于质谱的bottom-up蛋白质组学，在蛋白质酶切后用凝胶进行分级，级分的等电点和分子质量信息帮助鉴定蛋白质亚型[43]；热蛋白质组学（thermal proteome profiling，TPP）也可以用来表征蛋白质变体，TPP依靠bottom-up蛋白质组学方法来测量细胞的蛋白质热稳定性，不同蛋白质变体预期会有不同的热分布特征[44]，基于这种现象，Kurzawa等[45]对多个细胞系样品进行热处理，使用高分辨率等电聚焦分级样品肽段，以高肽段覆盖率测量热稳定性，将相似的肽段熔融图谱进行聚类，利用多肽水平的TPP数据和蛋白质变体检测算法检测出了不同剪接亚型，实现把肽段分配给特定的蛋白质变体，并对鉴定的蛋白质变体进行功能检测。

随着蛋白质检测技术的发展，越来越多的可变剪接蛋白质亚型被检测，通过检测蛋白质水平，剪接事件的生物学意义得到了更深的理解，为进一步阐明不同亚型的生物学功能奠定了基础。

3　利用蛋白质组学数据分析可变剪接

液相色谱-质谱/质谱（liquid chromatography-mass spectrometry/mass spectrometry，LC-MS/MS）运行后产生大量包含保留时间、*m*/*z*（mass-to-charge）值、强度等信息的谱图，一些搜索引擎如MaxQuant[46]、Mascot[47]、pFind[48]和MSFragger[49]等实现数据的处理和鉴定。其中，通过数据依赖采集（data dependent acquisition，DDA）方法采集的质谱数据通常采用数据库搜索的方法，即将实验获得的质谱数据与蛋白质数据库进行比对，来鉴定样品中的肽段和蛋白质，针对数据非依赖采集（data independent acquisition，DIA）解析的方法主要有谱库搜索方法、蛋白质序列库直接搜索方法、伪二级谱图鉴定方法等[50]。鉴定可变剪接蛋白亚型最常用数据库搜索的方法，把鉴定的肽段序列与数据库中的已知蛋白质序列比对鉴定蛋白质。

鉴定可变剪接蛋白亚型有几点显著问题：酶解产生的过小片段在分离过程中可能会被洗脱，部分达不到质谱的检测限，存在肽段覆盖率低的问题[51]；在数据分析角度，蛋白质的推断问题一直存在，一个基因的不同剪接蛋白亚型序列相似性往往很高，有的亚型间仅有部分氨基酸替换或者缺失；并且现有参考数据库包含剪接信息较少、有大量的共享肽段分配到蛋白质时存在歧义。目前，在数据处理过程中，改善可变剪接蛋白质变体鉴定困难的两种主要方法是优化蛋白质序列数据库和蛋白质的鉴定算法。

3.1　蛋白质序列数据库构建的研究进展

在进行蛋白质鉴定时，需要对照理论酶切肽段鉴定实验得到的肽段，其中参考序列数据库是蛋白质鉴定的基础。序列数据库的质量和完整性直接影响到搜索结果，为了确保能尽可能发现全面的异构体，数据库应该尽量包含所有可能的序列。同时，数据库与样本之间存在较大不一致会对搜索结果产生负面影响，不匹配的数据库可能会降低搜索的灵敏度和正确性或无法鉴定出需要的蛋白质。

3.1.1 公共参考蛋白质数据库

获取参考序列最简单的方式是直接使用公共数据库资源，常用的蛋白质数据库UniProtKB/SwissProt[52]提供可变剪接的注释信息，它的BLAST工具可以比较不同样本中的mRNA序列或者蛋白质序列，找出是否存在可变剪接的情况；RefSeq、EnSembl数据库提供剪接异构体的序列资源[53]，RefSeq提供非冗余的序列信息；国际人类蛋白质组计划的参考数据库NextProt[54]，可以提供表达数据、蛋白质组学数据和变异数据，截至2020年报道了一半的人类蛋白质组的额外剪接亚型（10 535个条目）[55]，数据库的工具特殊肽段检查器[56]同时考虑了同量异序取代、可变剪接和单氨基酸变体的情况，标示为Pseudo-Unique peptide（伪独特肽段）。

虽然这些数据库开始注意收录包括可变剪接在内的更全面的蛋白质信息，但使用公共数据库发现可变剪接蛋白质的数量还是有限的，一些可变剪接产生的新肽段无法鉴定。

3.1.2 转录组信息定制序列数据库

随着RNA测序技术进步和成本下降，使用转录组信息生成的定制蛋白质序列数据库的策略得到许多应用，通过测序数据补充公共数据库中缺失的肽段序列帮助识别可变剪接蛋白质，这种集成RNA测序数据和质谱数据的方式是蛋白质基因组学[57-58]的一部分。构建一个搜库的定制蛋白质序列数据库，首先获得转录本测序读段数据，将测序读段计数、映射发现剪接信息，这里的做法和转录组鉴定可变剪接事件一致；后续利用一帧、三帧或六帧的翻译技术把转录本翻译成蛋白质序列，最后将相应的蛋白质变异添加到数据库中。RNA数据可以直接对样本RNA进行测序[59]，也可以直接根据公共RNA序列数据库识别可变剪接[60]。

直接使用数据库测序数据，不对样品进行RNA测序。SpliceProt[60]是一个基于人类转录组实验数据鉴定出可变剪接再翻译构建的蛋白质序列数据库，RNA序列来源于NCBI网站的序列读取存档（SRA）存储库中获得表达序列标签（EST）、参考序列数据库（RefSeq）转录本和高通量测序（HTS）的数据，应用三元矩阵方法识别可变剪接，分析转录本的剪接位点，将信息存储在一个三种字符组成的矩阵中，通过聚类和比对算法来识别出不同的剪接模式和可变剪接事件，最后计算机翻译转录本为蛋白质序列。Da Silva等[35]利用SpliceProt方法鉴定可变剪接转录本，再用NMD分类器软件去除有提前终止密码子的变体，之后翻译序列用于MS数据搜库。同样利用公共RNA序列数据库信息，Lobas等[34]开发了新算法，样品没有进行RNA测序的情况下，对RefSeq数据库中的序列设置相应过滤条件，并开发模拟剪切过程的程序生成肽段构成数据库。

直接进行RNA测序生成样品定制序列数据库有更好的特异性（图1），Miller等[59]利用长读测序数据构建全长蛋白质数据库，Nextflow管道处理PacBio长读长测序数据，将全长转录本转换为蛋白质数据库，将PacBio数据和GENECODE数据库结合生成样品特异性数据库。Brandsma等[61]收集临床组织样品之后按照蛋白质基因组学的工作流程，根据RNA测序数据预测每个样品中存在的蛋白质序列变异，创建样品特异性蛋白质参考数据库用于肽和蛋白质的鉴定和定量，可以鉴定患者特异性非同义变异（包括剪接变异）和新的转录本亚型，比较疾病组和对照组的转录本表达差异分析疾病机制。

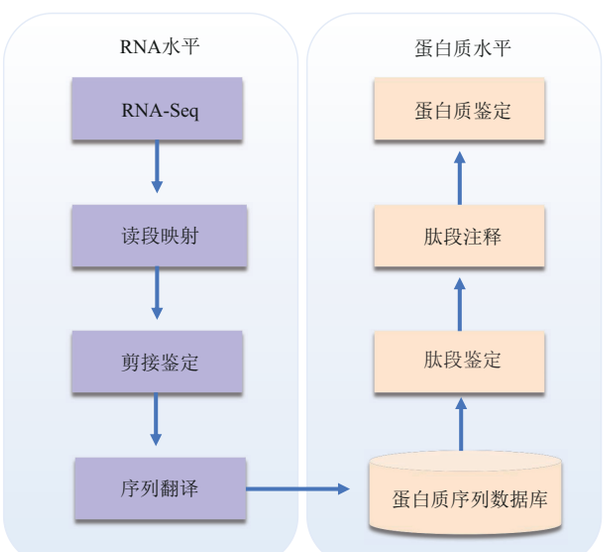


Fig. 1 Flowchart of constructing sample-specific databases to identify alternative splicing protein isoforms

图1 构建样本特异性数据库鉴定可变剪接蛋白质亚型流程

辅助建立序列数据库的一些工具，例如R包customProDB[62]可以将RNA-Seq数据生成蛋白质数据库，过程包含转录本的表达过滤器步骤，并且还可以注释相关变体，PASS[63]是一个用于筛选蛋

白质组中可变剪接蛋白质的工具，整合了从RNA-seq读段到建立蛋白质序列数据库、完成搜库的所有流程。

转录组测序针对所有mRNA进行测序，而全长翻译组测序（full-length translating mRNA sequencing，RNC-seq）和核糖体印迹测序（ribosomal profiling sequencing，Ribo-seq）针对翻译组[64]（正在翻译的mRNA）测序，收集的数据更接近产物蛋白质，所以被认为可以有效解决部分由转录本无法翻译成蛋白质所导致的假阳性问题。程序PROTEOFORMER[65]整合了从Ribo-seq数据到质谱搜库的过程，可以识别新的蛋白质变体。RNC-Seq相比Ribo-seq读段长并且容易排除干扰，Wu等[66]使用RNC-seq测序技术对整个翻译mRNA进行长读测序，翻译构建包含所有翻译剪切异构体的蛋白质数据库，搜库鉴定到的蛋白质数目比neXtProt数据库增加70%，具有唯一肽段的蛋白质变体更多。

这种通过转录本数据增加剪接蛋白质亚型的方法需要注意控制假阳性，在过程中对转录本进行一些质控筛选并且注意数据库的大小。Han等[67]利用靶向蛋白质组学和神经网络验证基于蛋白质基因组学方法得到的可变剪接蛋白质亚型；采用*De novo*测序后和数据库序列对照，Sinitcyn等[68]使用6种蛋白酶消化6种不同细胞系、深度分离结合3种串联质谱分析方法达到深度蛋白质测序，对比普通鸟枪法大大提高了序列覆盖率，产生足够数据支持进行从头组装蛋白质，证实大约64%剪接事件确实被翻译为相应蛋白质。

3.2　蛋白质可变剪接鉴定的算法进展

酶解后生成的肽段不再与其源蛋白质直接对应，鉴定出跨过剪接点的肽段的情况下才容易区分出不同的蛋白质亚型，并且大约50%的肽段会在两种或以上的蛋白质之间共享[69]，可变剪接蛋白质亚型间序列相似，更加剧了共享肽段的情况。由可变剪接等因素产生的截短蛋白质变体[70]，较短的亚型序列会包含在较长的亚型序列中，那么这种短亚型就很可能鉴定不出来，以及类似地缺乏独特肽段（unique peptide）的蛋白质家族成员之间也会区分困难。普通的鉴定算法[71]对无法区分的蛋白质按匹配到的肽段数目从多到少排序，由肽段子集构成的蛋白质往往被忽视。

针对可变剪接蛋白亚型的复杂鉴定问题已经开发了一些推断算法，结合现有搜索引擎更好地表征和检测可变剪接蛋白质，主要有肽段分组策略、肽段相关性理论、机器学习和剪接蛋白质有限表达理论几种方案（图2）。

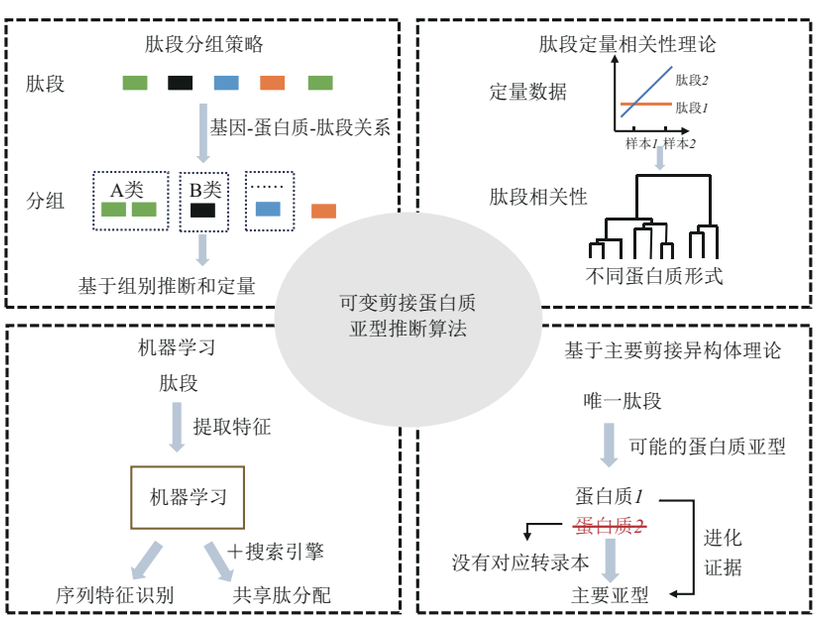


Fig. 2　Algorithm for inferring alternative splicing protein isoforms

图2　可变剪接蛋白质亚型推断算法

3.2.1 利用肽段分组策略推断剪接蛋白质亚型

研究者们提出依据蛋白质-肽段、基因-肽段和基因-蛋白质-肽段的对应关系进行肽段分组替代单个肽段作为分析对象，分组手段主要有直接命名分组和利用图论模型分组，分组后采用定义的肽段组别代替单个肽段进行蛋白质的鉴定或预测。

为了在肽段到蛋白质的推断中正确合理地分配共享肽段，有学者提出将肽段按照与蛋白质的对应关系进行分组[72]准确表征蛋白质，利用简化的组来解释在数据集中观察到的复杂的情况，使分析可以专注于样品中存在的蛋白质，简化数据解释。这种方案还可以计算实验中确定的蛋白质数量，帮助提高蛋白质定量的准确性。

2010年*Nature Biotechnology*刊登了一个用于蛋白质推断的肽段分类器（PeptideClassifier）[73]，基于基因模型（DNA特定区域）-蛋白质序列-蛋白质识别符（蛋白质ID）之间的关系，根据肽段相对于蛋白质序列和基因模型的信息含量将肽段分类。在这些分类中：1a类代表明确识别的某个蛋白质的序列；1b类也代表对应着唯一的蛋白质的一个独特序列，但其可能来自基因模型的不同剪接异构体转录本；2a类是识别一个蛋白质的子集肽段；2b类可以代表由一个基因编码的所有蛋白质的共有序列；3a类多肽明确地识别一个蛋白质序列，并且由来自不同基因座的多个基因模型编码；3b是来自不同基因模型编码的不同蛋白质序列。通过这样的分类方式把共享肽段进一步区分开，对于1a、1b和3a类信息足够明确，可以直接关联到一个特定的蛋白质序列，进一步区分不同的剪接异构体。

利用图论模型进行命名分类，Dou等[74]建立基于肽段-蛋白质-基因关系的三元图模型，成功识别了剪接蛋白质亚型，这种方案定义结构等效肽段（structurally equivalent PEPtides，SEPEPs）作为量化单位，SEPEPs表示在三元图中结构等价的一组多肽，根据SEPEPs与三部图中蛋白质和基因的连接模式共分为5类：C1至C5分别表示单异构型SEPEP，完全区别性SEPEPs，部分区别性SEPEPs，非区别性SEPEPs和多基因SEPEPs。依据肽段和基因的对应关系将肽段分为多基因对应肽段和单基因肽段，单基因肽段又可以根据与蛋白质的对应关系分为单异构体肽段（一种蛋白质对应的肽段）和多异构体肽段，其中多异构体肽段又可以分为完全区别肽段、部分区别肽段、非区别性肽段。将传统的肽段推断蛋白质的步骤变成SEPEPs的鉴定，之后在SEPEPs水平上去定量和鉴定不同的蛋白质异构体，更多地保留并利用有可能区分不同蛋白质亚型的肽段信息，值得注意的是，在数据集的测试中SEPEPs水平的定量比传统定量方法的基因覆盖率高5.8~33.8倍，进一步加强了对蛋白质亚型的表征。

以基因为中心的分类推理算法GpGrouper[75]用基因特异性多肽来指导共享多肽的分配，对蛋白质之间的共享肽段进行加权，改善了蛋白质定量的准确性。但是这种方法对肽段的序列覆盖度有一定要求，鉴定准确性有待评估，并且缺乏对蛋白质亚型的分析。

总的来说，这类方法充分利用所有可用的肽段信息，不至于忽视共享肽段的存在，让鉴定和定量更准确，对肽段进行分组帮助研究者更快地找到最适合自己研究目标的肽段。

3.2.2 利用肽段定量相关性理论鉴定剪接蛋白质亚型

蛋白质定量策略假设来自同一蛋白质的多个多肽在不同处理组中的定量表现相似，利用定量数据可以帮助解决蛋白质的鉴定问题，这种肽段相关性分析方法通过分析不同生物学环境中肽丰度的变化来鉴定蛋白质变体[76]。

PQPQ[77]的设计基于一个假设，即一种蛋白质的多肽的定量模式将在几个样本间相互关联，不同蛋白质的存在会导致肽段定量的不相关性，通过相关性分析检测不同的亚型；PeCorA[78]检测映射到同一蛋白质的多肽之间的数量差异，通过比较不同样品中特定肽段的丰度模式，找出反映不同蛋白质变体的差异丰度肽段，来检测有生物学意义的蛋白质变体变化；COPF[79]实现在肽段数据集中挖掘蛋白质变体组信息，利用数据集中不同蛋白质变体组的肽段强度数据差异、同一蛋白质变体组的肽段具有特定计量比，系统地将多肽归到正确的蛋白质变体组，直接从自下而上的蛋白质组数据集中检测蛋白质变体并对不同变体的功能含义进行系统评估，成功检测到两种选择性剪接的核蛋白，与之前的研究结果相符。

使用定量信息推断蛋白质，在定量不一致的肽段数据中发现不同蛋白质变体，这种方法需要注意的是它的敏感性很大程度上取决于定量方式和定量数据的质量。基于这种策略的工具包SpliceVista[80]根据提供定量质谱数据绘制每个肽段的定量模式，

提供基于其定量模式的肽段聚类选项。

3.2.3　利用机器学习推断剪接蛋白质亚型

基于机器学习方法可以帮助解决蛋白质的推断问题[81]，训练算法学习数据并做出预测或决策，自动化数据分析和应用智能决策提高工作效率和准确性，机器学习的主要步骤包括数据收集和预处理、特征提取、模型选择与训练、模型评估和模型应用。支持向量机（support vector machine，SVM）是一种机器学习方法，用于分类、回归和其他学习任务，工具IsoSVM[82]基于SVM分类器，通过对蛋白质序列进行特征提取和监督学习，自动识别异构体，达到区分异构体和同源基因蛋白质的效果。

也可以使用机器学习模型改进常用的搜索引擎以达到额外的鉴定效果，自动化算法Re-Fraction[83]利用凝胶电泳级分中的信息辅助推断蛋白质，构建基于凝胶电泳分级样品的蛋白质组学数据的支持向量机回归模型。首先从MaxQuant结果中筛选符合条件的肽段，设置一定规则将每个特殊肽段分配给最合适的组分，提取与级分关系确定的蛋白质的物理属性（质量、长度、肽段数量和等电点）特征，使用LibSVM支持向量机库的机器学习算法学习提取到的特征，最终该模型可以预测每个蛋白质相应的组分；面对某个肽段在几个蛋白质中共有的情况，利用模型预测的组分信息检查共享肽段是否在这些蛋白质中真实存在，级分中该共享肽段光谱计数为零的蛋白质会被删除，从而纠正共享肽的错误分配，提高唯一肽段的识别比例，应用该算法鉴定蛋白质组中的剪接变体和同源蛋白质效果好，实现了对现有鉴定算法的优化。

此外机器学习的方法经常用在RNA水平的剪接鉴定，例如使用RNA序列特征来预测可靶向的剪接缺陷[84]，利用机器学习对反式剪接因子的活动进行建模和预测[85]。

3.2.4　基于主要剪接异构体理论推断剪接蛋白质亚型

一个基因的可变剪接会产生不同转录本，由于表达水平不同会有优势异构体[86]，有学者提出大多数蛋白质编码基因都包含单一的显性异构体的理论[87]。关注样本中剪接蛋白质的主要亚型，ASV-ID方法[88]利用目标样本的基因表达数据帮助识别剪接变体，首先采用蛋白质基因组学方法构建序列数据库，肽段匹配蛋白质时去掉数据库搜索结果中没有RNA-Seq数据支持的异构体，这种方法找出更可靠的剪接蛋白质亚型，降低了蛋白质组的复杂性。

也有一些工作只关注主要亚型的鉴定，APPRIS数据库[89-90]注释基因的主要亚型和非主要亚型，主要亚型的选择基于一些进化证据，包括功能的保守性、结构基序的保守性和跨物种的保守性，这样选择是因为观察到大多数可变的亚型缺乏保守的结构或功能区域。但是主要剪接异构体的判定并没有一个标准，不同的方法有不同的选择结果，通过和蛋白质组得到的亚型数据对比验证主要亚型的一致性[91]。正如前面提到的，蛋白质的剪接亚型存在程度仍然存在争议，主要亚型的存在不代表其他非主要亚型的作用不重要，我们应该理性看待，进一步提高蛋白质水平的检测能力。

上述方法根据各自的策略实现了可变剪接蛋白质的鉴定，使用时根据实际情况尝试不同算法，但算法的性能受限于输入的数据质量，要保证可靠的鉴定，需要做好数据库和肽段鉴定的工作。

4　总结与展望

可变剪接是机体重要的调节机制，影响参与药物代谢、通路激活和细胞凋亡等相关基因的表达，同时，异常的剪接是重要的致病因素，越来越多的研究开始关注可变剪接和疾病的关联。因此，对可变剪接事件的检测具有重大意义，特别是对可变剪接蛋白质亚型的表征，有利于加深人类对生命过程和本质的理解。

总的来看，鉴定可变剪接的大部分工作在RNA水平，但大家已经开始关注蛋白质层面的鉴定，利用基于质谱的bottom-up蛋白质组学鉴定剪接体，可以从样品制备、分离和分析等过程入手解决，同时也一定要从数据分析角度考虑，它与分离技术、质谱技术相比同样是关键的步骤。然而，从bottom-up蛋白质组学数据中鉴定可变剪接蛋白质亚型的算法软件还是相对匮乏的，还没有哪种算法效果突出被广泛采用，大部分都需要配合搜库引擎一起使用。但是这些算法和策略不仅能用在可变剪接蛋白质异构体，还可以鉴定其他蛋白质变体，例如文中提到的蛋白质基因组学方法整合蛋白质组学、基因组学和转录组学，除了用于构建蛋白质序列数据库以识别剪接蛋白质亚型外，这种整合多组学的思路也被用于疾病的临床表征[92]等领域，促进精准医疗的发展。当然，构建数据库和改进鉴定算法这两种策略可以一起使用，以达到更好的效

果，另外在鉴定过程中应该注意对肽段鉴定的质量控制，保证鉴定的可靠性。

参 考 文 献

[1] Wang Y, Navin N E. Advances and applications of single-cell sequencing technologies. Mol Cell, 2015, **58**(4): 598-609

[2] Wright C J, Smith C W J, Jiggins C D. Alternative splicing as a source of phenotypic diversity. Nat Rev Genet, 2022, **23**(11): 697-710

[3] Xie J Q, Zhou X, Jia Z C, *et al*. Alternative splicing, an overlooked defense frontier of plants with respect to bacterial infection. J Agric Food Chem, 2023, **71**(45): 16883-16901

[4] Carbonara K, Andonovski M, Coorssen J R. Proteomes are of proteoforms: embracing the complexity. Proteomes, 2021, **9**(3): 38

[5] Smith L M, Kelleher N L, Proteomics C F T D. Proteoform: a single term describing protein complexity. Nat Methods, 2013, **10**(3): 186-187

[6] Rogalska M E, Vivori C, Valcárcel J. Regulation of pre-mRNA splicing: roles in physiology and disease, and therapeutic prospects. Nat Rev Genet, 2023, **24**(4): 251-269

[7] Wahl M C, Will C L, Lührmann R. The spliceosome: design principles of a dynamic RNP machine. Cell, 2009, **136**(4): 701-718

[8] Lee Y, Rio D C. Mechanisms and regulation of alternative pre-mRNA splicing. Annu Rev Biochem, 2015, **84**: 291-323

[9] Kastner B, Will C L, Stark H, *et al*. Structural insights into nuclear pre-mRNA splicing in higher eukaryotes. Cold Spring Harb Perspect Biol, 2019, **11**(11): a032417

[10] Wilkinson M E, Charenton C, Nagai K. RNA splicing by the spliceosome. Annu Rev Biochem, 2020, **89**: 359-388

[11] Nikom D, Zheng S. Alternative splicing in neurodegenerative disease and the promise of RNA therapies. Nat Rev Neurosci, 2023, **24**(8): 457-473

[12] Nagasaki H, Arita M, Nishizawa T, *et al*. Species-specific variation of alternative splicing and transcriptional initiation in six eukaryotes. Gene, 2005, **364**: 53-62

[13] Pan Q, Shai O, Lee L J, *et al*. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet, 2008, **40**(12): 1413-1415

[14] Mazin P V, Khaitovich P, Cardoso-Moreira M, *et al*. Alternative splicing during mammalian organ development. Nat Genet, 2021, **53**(6): 925-934

[15] Sciarrillo R, Wojtuszkiewicz A, Assaraf Y G, *et al*. The role of alternative splicing in cancer: from oncogenesis to drug resistance. Drug Resist Updat, 2020, **53**: 100728

[16] Boise L H, González-García M, Postema C E, *et al*. Bcl-x, a bcl-2-related gene that functions as a dominant regulator of apoptotic cell death. Cell, 1993, **74**(4): 597-608

[17] Zhang Y, Cai Q, Luo Y, *et al*. Integrated top-down and bottom-up proteomics mass spectrometry for the characterization of endogenous ribosomal protein heterogeneity. J Pharm Anal, 2023, **13**(1): 63-72

[18] Gabut M, Samavarchi-Tehrani P, Wang X, *et al*. An alternative splicing switch regulates embryonic stem cell pluripotency and reprogramming. Cell, 2011, **147**(1): 132-146

[19] Vecellio Reane D, Vallese F, Checchetto V, *et al*. A MICU1 splice variant confers high sensitivity to the mitochondrial Ca^{2+} uptake machinery of skeletal muscle. Mol Cell, 2016, **64**(4): 760-773

[20] Kjer-Hansen P, Weatheritt R J. The function of alternative splicing in the proteome: rewiring protein interactomes to put old functions into new contexts. Nat Struct Mol Biol, 2023, **30**(12): 1844-1856

[21] Biamonti G, Catillo M, Pignataro D, *et al*. The alternative splicing side of cancer. Semin Cell Dev Biol, 2014, **32**: 30-36

[22] Bonnal S, Vigevani L, Valcárcel J. The spliceosome as a target of novel antitumour drugs. Nat Rev Drug Discov, 2012, **11**(11): 847-859

[23] van Dijk E L, Jaszczyszyn Y, Naquin D, *et al*. The third revolution in sequencing technology. Trends Genet, 2018, **34**(9): 666-681

[24] Tay A P, Hamey J J, Martyn G E, *et al*. Identification of protein isoforms using reference databases built from long and short read RNA-sequencing. J Proteome Res, 2022, **21**(7): 1628-1639

[25] Hu T, Chitnis N, Monos D, *et al*. Next-generation sequencing technologies: an overview. Hum Immunol, 2021, **82**(11): 801-811

[26] Jiang W, Chen L. Alternative splicing: human disease and quantitative analysis from high-throughput sequencing. Comput Struct Biotechnol J, 2020, **19**: 183-195

[27] Zhang H, Jain C, Aluru S. A comprehensive evaluation of long read error correction methods. BMC Genomics, 2020, **21**(Suppl 6): 889

[28] Zhang G, Zhang Y, Jin J. The ultrafast and accurate mapping algorithm FANSe3: mapping a human whole-genome sequencing dataset within 30 minutes. Phenomics, 2021, **1**(1): 22-30

[29] Dobin A, Davis C A, Schlesinger F, *et al*. STAR: ultrafast universal RNA-seq aligner. Bioinformatics, 2013, **29**(1): 15-21

[30] Bryant D W, Shen R, Priest H D, *et al*. Supersplat-spliced RNA-seq alignment. Bioinformatics, 2010, **26**(12): 1500-1505

[31] Alser M, Rotman J, Deshpande D, *et al*. Technology dictates algorithms: recent developments in read alignment. Genome Biol, 2021, **22**(1): 249

[32] Patro R, Duggal G, Love M I, *et al*. Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods, 2017, **14**(4): 417-419

[33] Hsu M K, Lin H Y, Chen F C. NMD Classifier: a reliable and systematic classification tool for nonsense-mediated decay events. PLoS One, 2017, **12**(4): e0174798

[34] Lobas A A, Solovyeva E M, Levitsky L I, *et al*. Identification of alternative splicing in proteomes of human melanoma cell lines without RNA sequencing data. Int J Mol Sci, 2023, **24**(3): 2466

[35] da Silva E M G, Santos L G C, de Oliveira F S, *et al*. Proteogenomics reveals orthologous alternatively spliced proteoforms in the same human and mouse brain regions with differential abundance in an Alzheimer's disease mouse model. Cells, 2021, **10**(7): 1583

[36] Guo X, Zhao Y, Nguyen H, *et al*. Quantitative analysis of alternative pre-mRNA splicing in mouse brain sections using RNA

in situ hybridization assay. J Vis Exp, **2018**(138): 57889

[37] Yates J R. Mass spectrometry and the age of the proteome. J Mass Spectrom, 1998, **33**(1): 1-19

[38] Chen B, Brown K A, Lin Z, *et al*. Top-down proteomics: ready for prime time?. Anal Chem, 2018, **90**(1): 110-127

[39] 孙瑞祥, 罗兰, 迟浩, 等. “自顶向下(top-down)”的蛋白质组学——蛋白质变体的规模化鉴定. 生物化学与生物物理进展, 2015, **42**(2): 101-114
Sun R X, Luo L, Chi H, *et al*. Prog Biochem Biophys, 2015, **42**(2): 101-114

[40] Cai W, Tucholski T M, Gregorich Z R, *et al*. Top-down proteomics: technology advancements and applications to heart diseases. Expert Rev Proteomics, 2016, **13**(8): 717-730

[41] Toby T K, Fornelli L, Kelleher N L. Progress in top-down proteomics and the analysis of proteoforms. Annu Rev Anal Chem, 2016, **9**(1): 499-519

[42] Su T, Hollas M A R, Fellers R T, *et al*. Identification of splice variants and isoforms in transcriptomics and proteomics. Annu Rev Biomed Data Sci, 2023, **6**: 357-376

[43] Pedersen S K, Harry J L, Sebastian L, *et al*. Unseen proteome: mining below the tip of the iceberg to find low abundance and membrane proteins. J Proteome Res, 2003, **2**(3): 303-311

[44] Le Sueur C, Hammarén H M, Sridharan S, *et al*. Thermal proteome profiling: insights into protein modifications, associations, and functions. Curr Opin Chem Biol, 2022, **71**: 102225

[45] Kurzawa N, Leo I R, Stahl M, *et al*. Deep thermal profiling for detection of functional proteoform groups. Nat Chem Biol, 2023, **19**(8): 962-971

[46] Verheggen K, Raeder H, Berven FS, *et al*. Anatomy and evolution of database search engines-a central component of mass spectrometry based proteomic workflows. Mass Spectrom Rev, 2020, **39**(3): 292-306

[47] Perkins D N, Pappin D J, Creasy D M, *et al*. Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis, 1999, **20**(18): 3551-3567

[48] Li D, Fu Y, Sun R, *et al*. pFind: a novel database-searching software system for automated peptide and protein identification *via* tandem mass spectrometry. Bioinformatics, 2005, **21**(13): 3049-3050

[49] Kong A T, Leprevost F V, Avtonomov D M, *et al*. MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. Nat Methods, 2017, **14**(5): 513-520

[50] 侯鑫行, 周丕宇, 宫鹏云, 等. 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. 生物化学与生物物理进展, 2022, **49**(12): 2364-2386
Hou X H, Zhou P Y, Gong P Y, *et al*. Prog Biochem Biophys, 2022, **49**(12): 2364-2386

[51] Tsiatsiani L, Heck A J R. Proteomics beyond trypsin. FEBS J, 2015, **282**(14): 2612-2626

[52] Magrane M, UniProt Consortium. UniProt Knowledgebase: a hub of integrated protein data. Database, 2011, **2011**: bar009

[53] Sulakhe D, D'Souza M, Wang S, *et al*. Exploring the functional impact of alternative splicing on human protein isoforms using available annotation sources. Brief Bioinform, 2019, **20**(5): 1754-1768

[54] Duek P, Gateau A, Bairoch A, *et al*. Exploring the uncharacterized human proteome using neXtProt. J Proteome Res, 2018, **17**(12): 4211-4226

[55] Zahn-Zabal M, Lane L. What will neXtProt help us achieve in 2020 and beyond?. Expert Rev Proteomics, 2020, **17**(2): 95-98

[56] Schaeffer M, Gateau A, Teixeira D, *et al*. The neXtProt peptide uniqueness checker: a tool for the proteomics community. Bioinformatics, 2017, **33**(21): 3471-3472

[57] Sheynkman G M, Shortreed M R, Cesnik A J, *et al*. Proteogenomics: integrating next-generation sequencing and mass spectrometry to characterize human proteomic variation. Annu Rev Anal Chem, 2016, **9**(1): 521-545

[58] Nesvizhskii A I. Proteogenomics: concepts, applications and computational strategies. Nat Methods, 2014, **11**(11): 1114-1125

[59] Miller R M, Jordan B T, Mehlferber M M, *et al*. Enhanced protein isoform characterization through long-read proteogenomics. Genome Biol, 2022, **23**(1): 69

[60] Tavares R, de Miranda Scherer N, Pauletti B A, *et al*. SpliceProt: a protein sequence repository of predicted human splice variants. Proteomics, 2014, **14**(2/3): 181-185

[61] Brandsma C A, Guryev V, Timens W, *et al*. Integrated proteogenomic approach identifying a protein signature of COPD and a new splice variant of SORBS1. Thorax, 2020, **75**(2): 180-183

[62] Wang X, Zhang B. customProDB: an R package to generate customized protein databases from RNA-Seq data for proteomics search. Bioinformatics, 2013, **29**(24): 3235-3237

[63] Wu P, Pu L, Deng B, *et al*. PASS: a proteomics alternative splicing screening pipeline. Proteomics, 2019, **19**(13): e1900041

[64] Zhao J, Qin B, Nikolay R, *et al*. Translatomics: the global view of translation. Int J Mol Sci, 2019, **20**(1): 212

[65] Verbruggen S, Ndah E, Van Criekinge W, *et al*. PROTEOFORMER 2.0: further developments in the ribosome profiling-assisted proteogenomic hunt for new proteoforms. Mol Cell Proteomics, 2019, **18**(8 suppl 1): S126-S140

[66] Wu C, Lu X, Lu S, *et al*. Efficient detection of the alternative spliced human proteome using translatome sequencing. Front Mol Biosci, 2022, **9**: 895746

[67] Han Y, Wood S D, Wright J M, *et al*. Computation-assisted targeted proteomics of alternative splicing protein isoforms in the human heart. J Mol Cell Cardiol, 2021, **154**: 92-96

[68] Sinitcyn P, Richards A L, Weatheritt R J, *et al*. Global detection of human variants and isoforms by deep proteome sequencing. Nat Biotechnol, 2023, **41**(12): 1776-1786

[69] Meyer-Arendt K, Old W M, Houel S, *et al*. IsoformResolver: a peptide-centric algorithm for protein inference. J Proteome Res, 2011, **10**(7): 3060-3075

[70] DeBoever C, Tanigawa Y, Lindholm M E, *et al*. Medical relevance of protein-truncating variants across 337, 205 individuals in the

UK Biobank study. Nat Commun, 2018, **9**(1): 1612

[71] Tyanova S, Temu T, Cox J. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. Nat Protoc, 2016, **11**(12): 2301-2319

[72] Nesvizhskii A I, Aebersold R. Interpretation of shotgun proteomic data: the protein inference problem. Mol Cell Proteomics, 2005, **4**(10): 1419-1440

[73] Qeli E, Ahrens C H. PeptideClassifier for protein inference and targeted quantitative proteomics. Nat Biotechnol, 2010, **28**(7): 647-650

[74] Dou Y, Liu Y, Yi X, *et al*. SEPepQuant enhances the detection of possible isoform regulations in shotgun proteomics. Nat Commun, 2023, **14**(1): 5809

[75] Saltzman A B, Leng M, Bhatt B, *et al*. gpGrouper: a peptide grouping algorithm for gene-centric inference and quantitation of bottom-up proteomics data. Mol Cell Proteomics, 2018, **17**(11): 2270-2283

[76] Lukasse P N J, America A H P. Protein inference using peptide quantification patterns. J Proteome Res, 2014, **13**(7): 3191-3199

[77] Forshed J, Johansson H J, Pernemalm M, *et al*. Enhanced information output from shotgun proteomics data by protein quantification and peptide quality control (PQPQ). Mol Cell Proteomics, 2011, **10**(10): M111.010264

[78] Dermit M, Peters-Clarke T M, Shishkova E, *et al*. Peptide correlation analysis (PeCorA) reveals differential proteoform regulation. J Proteome Res, 2021, **20**(4): 1972-1980

[79] Bludau I, Frank M, Dörig C, *et al*. Systematic detection of functional proteoform groups from bottom-up proteomic datasets. Nat Commun, 2021, **12**(1): 3810

[80] Zhu Y, Hultin-Rosenberg L, Forshed J, *et al*. SpliceVista, a tool for splice variant identification and visualization in shotgun proteomics data. Mol Cell Proteomics, 2014, **13**(6): 1552-1562

[81] Zhao C, Liu D, Teng B, *et al*. BagReg: protein inference through machine learning. Comput Biol Chem, 2015, **57**: 12-20

[82] Spitzer M, Lorkowski S, Cullen P, *et al*. IsoSVM—distinguishing isoforms and paralogs on the protein level. BMC Bioinformatics, 2006, **7**: 110

[83] Yang P, Humphrey S J, Fazakerley D J, *et al*. Re-fraction: a machine learning approach for deterministic identification of protein homologues and splice variants in large-scale MS-based proteomics. J Proteome Res, 2012, **11**(5): 3035-3045

[84] Gao D, Morini E, Salani M, *et al*. A deep learning approach to identify gene targets of a therapeutic for human splicing disorders. Nat Commun, 2021, **12**(1): 3332

[85] Mao M, Hu Y, Yang Y, *et al*. Modeling and predicting the activities of trans-acting splicing factors with machine learning. Cell Syst, 2018, **7**(5): 510-520.e4

[86] Hu J, Boritz E, Wylie W, *et al*. Stochastic principles governing alternative splicing of RNA. PLoS Comput Biol, 2017, **13**(9): e1005761

[87] Gonzàlez-Porta M, Frankish A, Rung J, *et al*. Transcriptome analysis of human tissues and cell lines reveals one dominant transcript per gene. Genome Biol, 2013, **14**(7): R70

[88] Jeong S K, Kim C Y, Paik Y K. ASV-ID, a proteogenomic workflow to predict candidate protein isoforms on the basis of transcript evidence. J Proteome Res, 2018, **17**(12): 4235-4242

[89] Rodriguez J M, Maietta P, Ezkurdia I, *et al*. APPRIS: annotation of principal and alternative splice isoforms. Nucleic Acids Res, 2013, **41**(Database issue): D110-D117

[90] Rodriguez J M, Pozo F, Cerdán-Vélez D, *et al*. APPRIS: selecting functionally important isoforms. Nucleic Acids Res, 2022, **50**(D1): D54-D59

[91] Pozo F, Rodriguez J M, Martínez Gómez L, *et al*. APPRIS principal isoforms and MANE Select transcripts define reference splice variants. Bioinformatics, 2022, **38**(Suppl_2): ii89-ii94

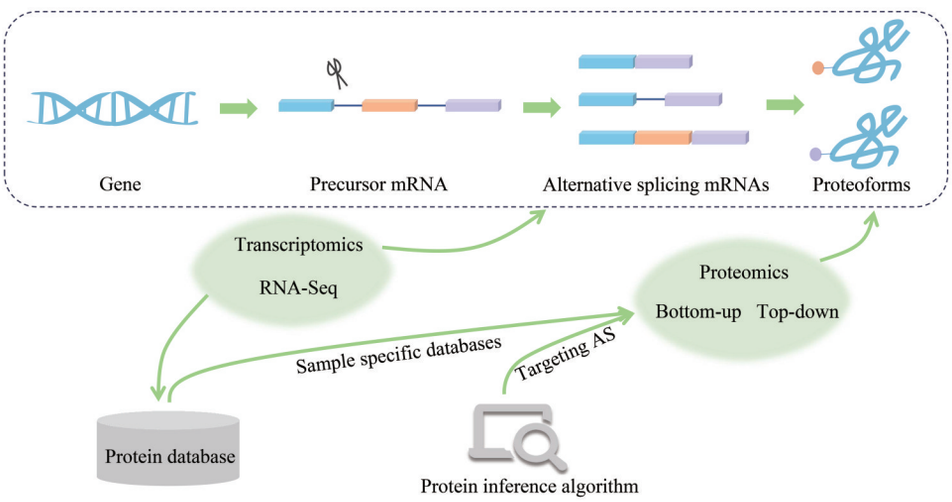
[92] Herbst S A, Vesterlund M, Helmboldt A J, *et al*. Proteogenomics refines the molecular classification of chronic lymphocytic leukemia. Nat Commun, 2022, **13**(1): 6226

Proteomics Data Reveals Alternative Splicing Proteoforms*

WU Yi-Ying, ZHANG Wei**, KONG De-Zhi**

(*Institute of Integrated Traditional Chinese and Western Medicine, Hebei Medical University, Shijiazhuang* 050017, China)

Graphical abstract



Abstract　Alternative splicing is an important regulatory mechanism in organisms, influencing the expression of genes involved in processes such as drug metabolism, pathway activation, and apoptosis. It refers to the process of removing introns from precursor mRNA and joining the remaining exons to produce mature mRNA. During this process, different combinations of exons can result in multiple mature mRNAs. This process is known as alternative splicing. Alternative splicing allows the same gene to produce different transcript variants and protein isoforms, increasing protein diversity and functional complexity. Transcriptomics and proteomics are two main approaches for identifying alternative splicing events. Transcriptomics identifies alternative splicing by analyzing differences between RNA sequencing data and reference sequences in databases. This method relies on the development of modern sequencing technologies. It also depends on increasingly improved splicing identification algorithms. Examples of these algorithms include alignment mapping and sequencing data quality control. The other approach is proteomic data analysis, which identifies corresponding protein products. We consider alternative splicing events more meaningful when they can be detected at the protein level. Alternative splicing proteoforms can be identified using bottom-up proteomics based on mass spectrometry. Due to the high sequence similarity between these alternative splicing proteoforms, general proteomic data analysis pipelines do not

* This work was supported by grants from The National Natural Science Foundation of China (82174004) and the Natural Science Foundation of Hebei Province (H2022206211, H2022206387).

** Corresponding author.

KONG De-Zhi. Tel：86-311-86266722，E-mail：kongdezhi@hebmu.edu.cn

ZHANG Wei. Tel：86-311-82266222，E-mail：weizhang@hebmu.edu.cn

Received：March 18，2024　Accepted：July 1，2024

achieve good discrimination between them. To improve the identification of proteoforms and obtain differentiation information for different isoforms in proteomic data, two strategies have been developed for improving data processing: the construction of special databases and targeted identification algorithms. We believe that this potential protein isoform information may play a crucial role in life science research. In terms of databases, it is not enough to only use ordinary public databases for searching. To ensure the discovery of as many isoforms as possible, the method of constructing sample-specific databases assisted by RNA sequencing data has been widely used, which can increase the probability of detecting proteoforms. Another key strategy is the improvement of protein identification algorithms. Traditional identification algorithms often struggle to distinguish between highly similar or mutually inclusive proteoforms. To address the complex identification of alternative splicing proteoforms, several inference algorithms have been developed, which are combined with existing search engines to better characterize and detect alternative splicing proteoforms. These include peptide grouping (PeptideClassifier, SEPepQuant, GpGrouper), peptide quantitative correlation (PQPQ, PeCorA, COPF, SpliceVista), machine learning (IsoSVM, Re-Fraction, LibSVM), and major splice isoform theory (ASV-ID). Such methods have shown promising results in focusing on alternative splicing proteoforms. When using these algorithms, we should try different ones based on actual situations. Additionally, the performance of these algorithms is limited by the quality of input data. To ensure reliable identification, it is also essential to perform proper peptide identification and quality control at the front end. In general, the detection and differentiation of spliced protein isoforms are still inadequate, requiring continued attention. This article reviews recent research progress on alternative splicing and its biological functions, as well as the detection of alternative splicing at different levels, and introduces the main methods for identifying alternative splicing proteoforms using bottom-up proteomic data. Identifying different alternative splicing proteoforms helps us understand the comprehensive functions of proteins and is of great significance for discovering related biomarkers and key drug targets.

Key words alternative splicing, mass spectrometry data analysis, protein identification algorithm, protein sequence database

DOI：10.16476/j.pibb.2024.0109