PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2023,50(1):109~125
www.pibb.ac.cn

蛋白质组学肽段鉴定可信度评价方法*

周文婧[1,2] 曾文锋[1] 迟 浩[1,2)**] 贺思敏[1,2)**]
([1] 中国科学院智能信息处理重点实验室，中国科学院计算技术研究所，北京 100190；[2] 中国科学院大学，北京 100049)

摘要 蛋白质组学基于质谱数据鉴定肽段和蛋白质，从而给出基因表达的直接证据，帮助解析蛋白质的结构和功能，研究蛋白质与疾病的关系，提供靶向治疗方案，而这些都取决于鉴定的肽段和蛋白质的准确性。蛋白质组学常采用目标-诱饵库方法（target-decoy approach，TDA）对鉴定的肽段和蛋白质进行质量控制，并对其进行改进演化后应用到子类肽段（比如突变肽段和修饰肽段等）和交联肽段等特殊鉴定结果的可信度评价中。然而，TDA存在两个局限，即错误率估计值不够准确以及不能评价单个鉴定结果的可信度，经过TDA质量控制后的结果还需要进一步检验，因此领域内也提出了一系列其他方法（本文统称为Beyond-TDA方法），协同加强肽段的可信度评价。本文对数据依赖模式下采集的质谱数据肽段层面的TDA常规方法和特殊方法进行了综述，对Beyond-TDA方法进行了分类阐述，并总结了各种方法的优势与不足。

关键词 蛋白质组学，质谱分析，目标-诱饵库方法，假发现率，可信度评价方法
中图分类号 Q51，TP39 **DOI**：10.16476/j.pibb.2022.0004

1 蛋白质组学与质谱分析

蛋白质组学以特定时空下的一组蛋白质为对象来研究基因和细胞的功能，质谱分析是蛋白质组学的常用手段[1]。在常规的自底向上的蛋白质组学中，生物样品中的蛋白质首先酶切为肽段，经过色谱分离后进入质谱，进行质量分析和检测，得到一级谱图。随后，质谱仪会从一级谱图中选取高丰度肽段信号进行碎裂，并采集二级谱图。一级谱图和二级谱图构成了串联质谱数据，其包含三维信息：肽段离子的质荷比、强度和保留时间。质谱分析是指从串联质谱数据中解析出生物样品包含的肽段和蛋白质。

质谱数据的解析结果对蛋白质组学研究至关重要。质谱数据中鉴定的肽段可以作为蛋白质存在的直接证据，进而证明基因表达活动[2-4]；同时，鉴定的肽段，特别是交联肽段，能够帮助解析蛋白质的结构，研究蛋白质的相互作用关系[4-6]；更重要的是，作为基因的直接表达产物，蛋白质含量的上下波动可以帮助发现致病基因及研制具有相应靶向作用的药物[7-9]。常用的质谱数据解析方法有数据库搜索[10-12]、肽段从头测序[13-16]和谱库搜索[17-18]等。得到质谱数据的初步解析结果后，需要对谱图和肽段层次的解析结果进行质量控制，即控制解析结果的错误率。这一过程也被称为过滤，即通过控制鉴定结果的错误率范围，过滤掉不可信鉴定结果，最终报告出可信结果。经过谱图和肽段层面的质量控制后，可以基于可信肽段推断蛋白质并进行蛋白质层面的质量控制，最终得到高可信蛋白质并进行下游生物学研究[19-22]。

然而，在目前的蛋白质组学研究中，质谱数据鉴定到的肽段和蛋白质的可信度可能仍然存在较大问题。造成错误鉴定的原因繁多，数据库不完整，单核苷酸突变，酶切位点、电荷、修饰类型、修饰位点的错误判断以及同位素峰的误匹配都可能造成错误鉴定[23-24]。如果对鉴定的肽段和蛋白质不进行严格的质量控制，会严重影响鉴定结果的可信度。2014年Kim等[2]和Wilhelm等[3]在《自然》

* 国家重点研发计划重点专项（2016YFA0501300）和国家自然科学基金优秀青年科学基金（32022046）资助项目。
** 通讯联系人。
迟浩 Tel：010-62600822，E-mail：chihao@ict.ac.cn
贺思敏 Tel：010-62600822，E-mail：smhe@ict.ac.cn
收稿日期：2022-01-06，接受日期：2022-03-23

（*Nature*）杂志同期发表了两项人类蛋白质组草图研究结果，是人类蛋白质组研究的里程碑。两篇文章均构建和使用了自定义的质谱数据解析流程，分别鉴定得到17 294和18 097个人类基因，覆盖了人类基因组的84%和92%。然而，两篇草图文章的质谱数据和鉴定结果公开后，领域对草图文章鉴定结果的可信度产生了质疑[25-27]。首先，人类蛋白质组草图研究中蛋白质的推断标准不严格，仅由单肽段鉴定的蛋白质也被保留，如果不考虑这部分结果，那么Kim等的文章会有5 288个基因被排除，而Wilhelm文章中也有1 259个仅由单肽段鉴定的蛋白质不能计入最终鉴定结果（未提供基因数目）[26]。另外，鉴定结果的准确度和灵敏度都存在问题。最为明显的错误是，两篇人类蛋白质草图文章都未制备嗅觉组织样品，但分别鉴定到了108个和200个嗅觉组织所特有的嗅觉受体蛋白质[25]，而嗅觉受体蛋白是一种跨膜蛋白，只能在鼻黏膜组织中才能鉴定到[28]。此外，本应普遍出现的3种细胞受体因子的表达模式没有在草图中得到鉴定，说明草图还远未达到完整[27]。低可信度的鉴定结果会影响后续对蛋白质结构、功能、相互作用关系和致病机理等的研究，所以对蛋白质组学质谱数据鉴定结果进行可信度评价极为关键。肽段的可信度是蛋白质可信度评价方法的前提和基础，领域内对于肽段的可信度评价方法研究更久更成熟，所以本文将重点对肽段的可信度评价方法进行综述。

肽段鉴定可信度评价方法历经了多次发展，早期主要使用基于阈值的评价方法，包括设定搜索引擎打分阈值、*P*-value和*E*-value等。设定搜索引擎打分阈值的方法是指对于搜索引擎给出的所有鉴定结果，将打分高于某特定阈值的结果认为是可信鉴定结果，打分低于特定阈值的结果认为是不可信鉴定结果[29-30]，比如有研究认为Mascot引擎打分超过30分的结果为可信鉴定结果[30]。这种设定打分阈值的方法使用简便，但是打分阈值的设定极大依赖于人工经验。*P*-value(x)是指在给定谱图的情况下，随机匹配打分大于x的概率[31]，*E*-value(x)是指在给定谱图和数据库的情况下，随机匹配打分大于x的肽段数目的期望[31]。这两者的关系为*E*-value(x)=n×*P*-value(x)，其中n为候选肽段数目。*P*-value和*E*-value让不同搜索引擎的鉴定结果的可信度变得可比，但是和打分阈值方法一样，*P*-value和*E*-value的阈值同样也依靠人工经验。

2002年，Keller和Nesvizhskii等[32]提出了基于贝叶斯公式的质量控制方法PeptideProphet，将概率模型引入肽段可信度评价方法。PeptideProphet方法认为正确肽段的打分服从高斯分布，错误肽段的打分服从伽马分布，并且对特异酶切位点数目不同和电荷数目不同的肽段分别拟合分布，估算每个肽段-谱图匹配是正确匹配的概率。为了适应不同的数据和实验，可以在以上分布的基础上，采用期望最大化方法（expectation maximization，EM）构建混合模型，不断迭代拟合正确和错误鉴定结果的分布。后续10年间，PeptideProphet衍生出了一系列方法。2003年，Nesvizhskii等[33]在PeptideProphet的基础上提出了评价蛋白质可信度的ProteinProphet方法，该方法认为蛋白质存在的概率可以通过该蛋白质鉴定的肽段至少有一条是正确的概率来估算。2007年，该团队提出了基于PeptideProphet的半监督模型[34]，将部分诱饵库鉴定结果用于EM训练中。随后，该团队提出了可变成分混合模型和半参数混合模型两种方法[35]，打破了PeptideProphet混合模型中限制参数估计的假设。2008年，该团队提出生成模型方法[36]，首先对谱图进行聚类，每一类估计一个混合模型。同时，对每张谱图的前10名候选肽段均计算PeptideProphet概率，并根据概率重新排列这些候选肽段的顺序。2011年，为了能够利用多种搜索引擎的特性，鉴定更多和更可信的肽段和蛋白质，该团队提出iProphet[37]，在PeptideProphet的基础上，结合重复实验鉴定情况、重复引擎鉴定情况、重复谱图、重复母离子和重复修饰等特征，能够合并多种搜索引擎和多次重复实验的结果，得到更好的混合模型。PeptideProphet方法经历了长久发展，在标注数据集上能取得较好的拟合效果，但该方法依赖于估计的数据分布与真实数据分布的相似程度，而且EM方法可能需要耗费较多的训练轮次和训练时间。

2007年，Elias和Gygi[19]总结并评测了Moore等提出的目标-诱饵库方法（target-decoy approach，TDA）[38-40]，通过估计假发现率（false discovery rate，*FDR*），对鉴定的肽段的可信度进行评价。*FDR*是对真实错误率的一种估计，通常只将*FDR*小于等于1%的鉴定结果作为可信结果。由于TDA方法公式简单、使用简便，它逐渐成为质谱数据解析过程中最主流的质量控制方法，并在子类肽段（包括一般子类肽段、突变肽段和修饰肽段等）和交联肽段等特殊鉴定目标的可信度评价中进行了衍

生和演化。本文将在第二节中重点讲述TDA常规方法及其特殊演化方法在蛋白质组学肽段鉴定可信度评价中的应用。

本文综述了蛋白质组学质谱数据鉴定的肽段的可信度评价方法。第一节讲述蛋白质组学质谱数据制备及数据分析方法，同时对质谱数据鉴定结果的可信度问题以及早期的肽段鉴定可信度评价方法进行阐述。第二节首先讲述评价肽段可信度的TDA常规方法，然后讲述在子类肽段和交联肽段等特殊鉴定目标中的TDA演化方法，最后讲述TDA方法的局限。第三节首先介绍肽段可信度评价方法的统一衡量指标——检验假阳率和检验假阴率，然后综述领域内现有的Beyond-TDA方法，即在TDA方法的基础上，对鉴定结果的可信度进行进一步检验，并对它们的检验假阳率和检验假阴率进行比较。第四节对全文内容进行总结。

2　目标–诱饵库方法（TDA）

随着质谱采集技术的快速进步和鉴定软件的蓬勃发展，一次质谱实验分析即可获取海量的肽段-谱图匹配结果，这些鉴定结果的准确性对后续生物分析至关重要。TDA（图1）可以实现对鉴定结果可信度的快速和相对准确地评估。本节将对TDA常规方法、特殊方法以及TDA方法的局限性进行详细阐述。

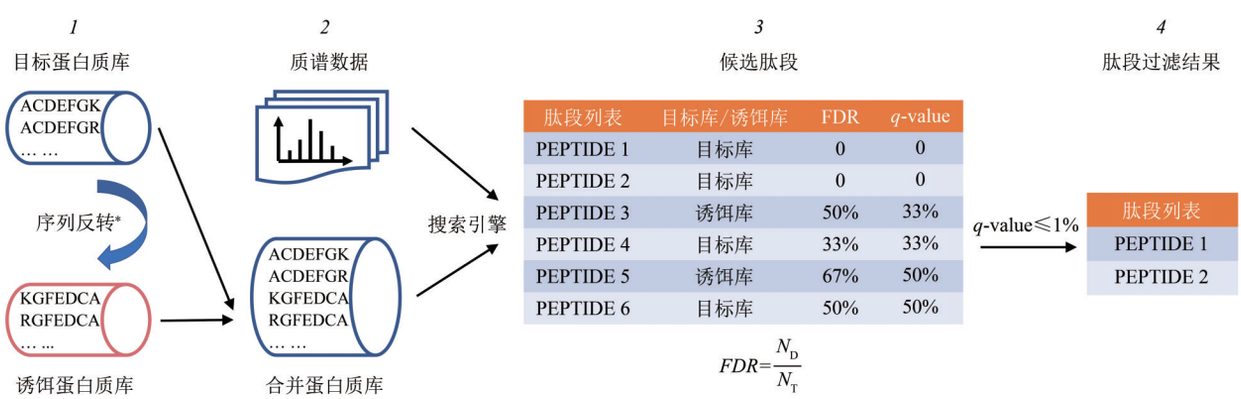


Fig. 1　Target–decoy approach

图1　目标–诱饵库方法

目标-诱饵库方法：*1*：目标蛋白质库序列反转得到诱饵蛋白质库；*2*：质谱数据搜索两个库的合并库；*3*：根据鉴定到的诱饵库肽段数目与目标库肽段数目的比值计算肽段层面的假发现率（*FDR*），N_T代表目标库鉴定结果数目，N_D代表诱饵库鉴定结果数目；*4*：采用假发现率的*q*-value小于等于1%过滤得到最终的肽段列表。*除了蛋白质序列反转，还可以通过肽段反转、氨基酸置换和马尔可夫方法构建诱饵蛋白质，示意图中只绘制一种方法。

2.1　常规方法

TDA方法通过构造诱饵蛋白质数据库（以下简称“诱饵库”）对鉴定结果进行质量控制。诱饵库的构建方式主要有4种：蛋白质序列反转[38-39]、肽段序列反转[19]、氨基酸随机置换[19]和马尔可夫方法[41]。蛋白质序列反转是将目标蛋白质数据库（以下简称“目标库”）的每个蛋白质序列整体进行N-C端方向反转，肽段反转是指将目标库蛋白质理论酶切后生成的所有肽段序列反转，随机置换是指将目标库蛋白质理论酶切后生成的所有肽段序列中的每个氨基酸与序列中的其他氨基酸的位置进行随机置换，马尔可夫方法是使用马尔可夫链从目标库学习到氨基酸分布规律，然后根据氨基酸分布规律构建诱饵库。前两种方法本质都是序列反转，后两种方法本质都是序列随机化。这4种方法均是为了构造与目标库同规模且同氨基酸分布的诱饵库。其中，蛋白质反转的方法最为常用。有研究表明，诱饵库构建方法对最终结果没有显著影响[42-43]，但是可以通过随机置换的方法生成多种随机库分别估计*FDR*后取平均值作为最终的*FDR*估计值，这样估计的*FDR*更接近真实错误率[44-46]。

TDA方法应用的前提是假设一次错误匹配结果（Elias和Gygi的文章描述为incorrect result，具体是指错误匹配中的随机匹配）匹配到目标库和诱饵库的概率是相等的。在此基础上，该假设通过匹配到的诱饵库鉴定结果的数目N_D来估计目标库鉴

定结果中的错误鉴定结果数目，用目标库错误鉴定结果数目比上所有的目标库鉴定结果数目N_T，就可以计算出目标库鉴定结果中的假发现率（*FDR*）：

$$FDR = \frac{N_D}{N_T} \quad (1)$$

TDA假设简单，实现方便，而且能对鉴定结果的可信度做出简单评估，具有相对合理性，比如*FDR*越小，过滤时的打分阈值越高，鉴定结果越可信。由于*FDR*并不随着鉴定结果打分的降低而单调递增，在实际实验中可能会出现鉴定结果高打分区域的*FDR*高于低打分区域的*FDR*，这样会影响根据*FDR*阈值进行过滤的实际操作。为了解决这个问题，在实际应用中通常使用*q*-value来替代*FDR*。*q*-value是指能过滤出打分为*x*的肽谱匹配结果所需要的*FDR*阈值的最小值[47]，相当于对*FDR*做了平滑操作，后续提到的*FDR*均指*q*-value。本文将采用TDA估计*FDR*进而对鉴定结果进行质量控制的方法称为“TDA-FDR”方法。

由于前述TDA-FDR方法不能评估单个鉴定结果的后验错误概率（posterior error probability，PEP），Local FDR方法逐渐得到发展和应用[34，48-49]。Local *FDR*是指打分等于*x*的鉴定结果中诱饵库鉴定结果和目标库鉴定结果的比例，而前述*FDR*是指全局*FDR*，即打分大于等于*x*的鉴定结果中诱饵库鉴定结果和目标库鉴定结果的比例。Kall等[48]的研究认为，在统计学意义上，Local *FDR*比*FDR*和*q*-value更保守。

质谱分析会给出每张谱图所对应的肽段信息，每个鉴定结果就是一个肽段-谱图匹配（peptide-spectrum match，PSM），由PSM可以得到肽段，而由肽段又可以推断出鉴定到的蛋白质，所以质谱鉴定结果包含谱图、肽段和蛋白质3个层面的鉴定信息。相应地，谱图、肽段和蛋白质3个层面均可估计各自的*FDR*。这3个层面的*FDR*估计基本方法均是通过当前打分阈值下的诱饵库鉴定结果（谱图/肽段/蛋白质）数目除以目标库鉴定结果数目。人类蛋白质组计划（Human Proteome Project，HPP）要求质谱分析中谱图、肽段和蛋白质3个层面的*FDR*均不能超过1%[50-51]。

2.2 特殊方法

TDA-FDR方法萌发于常规蛋白质组学，但蛋白质组学分析中常常会对某些特殊的鉴定结果感兴趣，比如子类肽段和交联肽段等，常规的TDA-FDR方法并不能直接用于特殊鉴定结果的可信度评价，需要针对特殊目标进行改进和演化。

2.2.1 针对子类肽段的TDA-FDR方法

对于某些子类鉴定结果，比如蛋白质基因组学分析在注释相对完全的物种中鉴定到的新肽段，或者富含翻译后修饰的鉴定结果，由于这些子类鉴定结果的数目相对于总的鉴定结果而言并不多，而这些子类肽段的搜索空间比常规肽段的搜索空间更大[52]，如果所有鉴定结果合并进行过滤会导致子类鉴定结果的*FDR*估计不准确[4，52-54]。所以，需要对每种子类鉴定结果单独计算*FDR*，即分开过滤，这种方法被称为“Separate FDR”，核心思想是对于鉴定结果按数据类型分组（鉴定到不同种类的翻译后修饰或者鉴定为新肽段或已注释肽段），在每组数据上单独使用TDA来估计组内数据的*FDR*并对组内数据进行过滤。Separate FDR方法计算公式如下：

$$FDR_k = \frac{N_{D\_k}}{N_{T\_k}} \quad (2)$$

其中*k*代表肽段类别，FDR_k代表第*k*类肽段的*FDR*，N_{D_k}代表第*k*类诱饵库肽段鉴定数目，N_{T_k}代表第*k*类目标库肽段鉴定数目。这种方法可以更准确地估计每类肽段的*FDR*，但是对于子类肽段数目比较敏感。当子类肽段数目较少时，计算的*FDR*可能并不准确。

李婧等[55]发现，对于突变肽段这种子类鉴定结果，即使采用Separate FDR方法，也不能有效解决突变肽段打分向低分区域聚拢的问题（即鉴定到的突变肽段不可信），她们认为子类数据中鉴定到的诱饵库结果可能与该子类数据占总体数据的比例有关，所以根据鉴定结果中的子类数据与总体数据的比例重新估计子类数据中的诱饵库鉴定结果数目，在此基础上重新估计子类数据的*FDR*。由于该方法最早用于估计突变肽段的*FDR*，所以称该方法为“Variant FDR”，计算公式如下：

$$FDR_{k^+} = \frac{N_{D^+} \times \frac{N_{D^-\_k}}{N_{D^-}}}{N_{T^+\_k}} \quad (3)$$

其中*k*代表肽段类别，FDR_{k^+}代表打分阈值之上的第*k*类肽段的*FDR*，N_{D^+}代表打分阈值之上的所有诱饵库肽段数目，N_{D^-}代表打分阈值之下的所有诱饵库肽段数目，$N_{D^-_k}$代表打分阈值之下的第*k*类诱饵库肽段数目，$N_{T^+_k}$代表打分阈值之上的第*k*类目标库肽段数目。基因组证据表明，Variant FDR方法比常规TDA-FDR和Separate FDR过滤出的突变

肽段的准确性更高。

当子类鉴定结果样本量较小时，即使是分开过滤，直接使用TDA公式计算得到的*FDR*可能并不准确，此时可以使用Transfer FDR方法估计任意数目的子类鉴定结果的*FDR*。该方法由付岩等[56]提出，通过线性拟合诱饵匹配中子类肽段比例与打分间的函数关系，更准确地估计打分阈值处的子类错误目标匹配数量，以此估计子类数据的*FDR*，避免子类数据样本数目较少带来的*FDR*估计不准确的问题。Transfer FDR的计算公式如下：

$$FDR_k = \frac{N(x)}{N_k(x)}(ax + b)FDR \tag{4}$$

其中k为肽段类别，FDR_k为第k类肽段的*FDR*，x代表肽段打分，$N(x)$代表打分超过x的所有肽段数目，$N_k(x)$代表打分超过x的第k类肽段数目，a和b代表线性拟合常数项，*FDR*代表所有肽段的全局*FDR*。

分开过滤的思想可以很自然地应用于蛋白质基因组学鉴定的新肽段和已注释肽段的可信度评价中。蛋白质基因组学是通过蛋白质组学鉴定蛋白质，结合基因组信息对生物的基因进行重注释，即发现新基因、新现象（比如新N端、可变剪接）和校正已注释基因，对应到质谱分析中主要为发现新肽段和校正已注释肽段[4, 54, 57]。Krug等[58]研究表明，对于大肠杆菌等注释程度较高的物种，鉴定到的新肽段的后验错误概率分布与诱饵库肽段的后验概率分布几乎相同，所以蛋白质基因组学发现新肽段需要进行严格质控。如果对新肽段单独估计*FDR*，可能会因为注释程度较高的物种中新肽段数目较少而导致估计值不够准确；而如果对新肽段和已注释肽段统一进行*FDR*估计，则会降低新肽段的准确度。Zhang等[54]将分开过滤的思想应用到蛋白质基因组学中，推导了已注释肽段和新肽段的*FDR*与全局*FDR*的关系，并证明了已注释肽段的*FDR*小于全局*FDR*小于新肽段*FDR*[59]，这两类肽段的*FDR*计算公式如下所示：

$$FDR_{new}(x) = \frac{FDR(x)}{FDR(x) + \frac{6(1-\theta)}{6-\mu}(1 - FDR(x))} \tag{5}$$

$$FDR_{ann}(x) = \frac{FDR(x)}{FDR(x) + \frac{6\theta}{\mu}(1 - FDR(x))} \tag{6}$$

公式（5）和公式（6）中，$FDR_{new}(x)$和$FDR_{ann}(x)$分别代表打分高于x的新肽段和已注释肽段的*FDR*，μ指基因组序列注释比例，θ指基因注释完整性比例。基因组序列注释比例是指已注释基因总长占基因组长度的比值，基因注释完整性比例是指已注释基因占基因组上所有真实表达基因的长度比例。这两个变量中，μ可以直接计算得到，但很遗憾的是，θ是未知量，无法得知，所以无法通过以上公式精确计算两类肽段的*FDR*。但是通过μ与θ的关系（由定义可知，$\mu \leqslant \theta$且$\theta > 0$）可以从公式（5）和公式（6）中推导出$FDR_{new}(x) > FDR(x) > FDR_{ann}(x)$[59]。

为了更精准地计算出两类肽段的*FDR*，张昆[59]又将Transfer FDR方法应用到蛋白质基因组学中，通过线性拟合的方法重新估计新肽段中的错误鉴定数目，单独计算酿酒酵母蛋白质基因组学分析中鉴定到新肽段的*FDR*（方法同公式（4））。酵母新肽段中的合成实验表明，蛋白质基因组学中，Transfer FDR方法比Separate FDR方法估计的新肽段*FDR*更准确。

2.2.2　针对交联肽段的TDA-FDR方法

交联蛋白质组学（这里特指二肽交联）质谱数据由两条相互交联的肽段碎裂打谱得到，与常规蛋白质组学鉴定肽段结果非对即错相比，交联蛋白质组学鉴定得到的是两条相互交联的肽段，它们存在全对、全错、一对一错这3种情况，这使得*FDR*的计算方式变为

$$FDR = \frac{N_{TD} - N_{DD}}{N_{TT}} \tag{7}$$

其中，N_{TD}代表交联肽段一条来自目标库，而另一条来自诱饵库的鉴定结果数目，N_{DD}代表交联肽段中两条肽段均来自诱饵库的鉴定结果数目，N_{TT}代表交联肽段中两条肽段均来自目标库的鉴定结果数目[6, 60]。在实际应用中，对于不同蛋白质之间（inter-protein）和同一蛋白质之内（intra-protein）这两类交联肽段要应用公式（7）分别计算*FDR*，这里也应用了分开过滤的思想[60-61]。

糖基化修饰是一种特殊的修饰，糖蛋白质组学中鉴定的糖肽可以看作是特殊的修饰肽段，但由于糖链的特殊性，不妨将糖肽看作糖链和肽段的交联，类似于交联蛋白质组学中的交联二肽。早期计算糖肽鉴定结果的*FDR*比较困难，因为难以对糖链构建诱饵库，所以无法直接估计糖链的*FDR*，仅通过估计肽段的*FDR*进行质控。2013年，Strum等[62]提出糖不变而蛋白质随机置换以及糖增加11 u而蛋白质不变两种方法构建诱饵库，这种方法

最早提出了诱饵糖库的思想，但只是改进打分，未对*FDR*进行研究。2017年，Liu等[63]提出将糖库中的理论Y离子质量随机增加1~30 u来构造糖链的诱饵谱图，作者想出该方法是受到了肽段诱饵库的启发，肽段诱饵库可以通过反转序列后生成诱饵谱图，也可以先生成谱图，然后谱峰偏移构建诱饵谱图，所以作者认为通过偏移糖库中的理论Y离子质量也可以达到构建糖链诱饵谱图的效果。通过鉴定的糖链诱饵谱图和肽段诱饵谱图的数目分别估计出糖链和肽段的*FDR*，然后用容斥原理估计出糖肽的*FDR*：

$$FDR(x) = FDR_{G}(x) + FDR_{P}(x) - FDR_{G\cap P}(x) \quad (8)$$

其中$FDR(x)$建模了糖肽鉴定错误的概率，$FDR_{G}(x)$建模了糖链鉴定错误的概率，$FDR_{P}(x)$建模了肽段鉴定错误的概率，$FDR_{G\cap P}(x)$建模了糖链和肽段同时鉴定错误的概率。

公式(7)与公式(8)形式上似乎差异很大，但实际上只是同种计算方法的不同呈现形式。前面提到可以将糖肽看作是糖链与肽段的交联，同时假设糖链和肽段各自的诱饵库鉴定结果数目与错误鉴定结果数目的比例是相等的，那么如果以N_{TD}代表糖链和肽段其中一个来自目标库而另一个来自诱饵库的鉴定结果数目，N'_{TD}代表糖链来自目标库而肽段来自诱饵库的鉴定结果数目，N'_{DT}代表糖链来自诱饵库而肽段来自目标库的鉴定结果数目，N_{DD}代表糖链和肽段均来自诱饵库的鉴定结果数目，N_{TT}代表糖链和肽段均来自目标库的鉴定结果数目，可知有$N_{TD} = N'_{TD} + N'_{DT}$，那么公式(8)可以变为：

$$FDR(x) = FDR_{G}(x) + FDR_{P}(x) - FDR_{G\cap P}(x) = \frac{N'_{DT}}{N_{TT}} + \frac{N'_{TD}}{N_{TT}} - \frac{N_{DD}}{N_{TT}} = \frac{N_{TD} - N_{DD}}{N_{TT}} \quad (9)$$

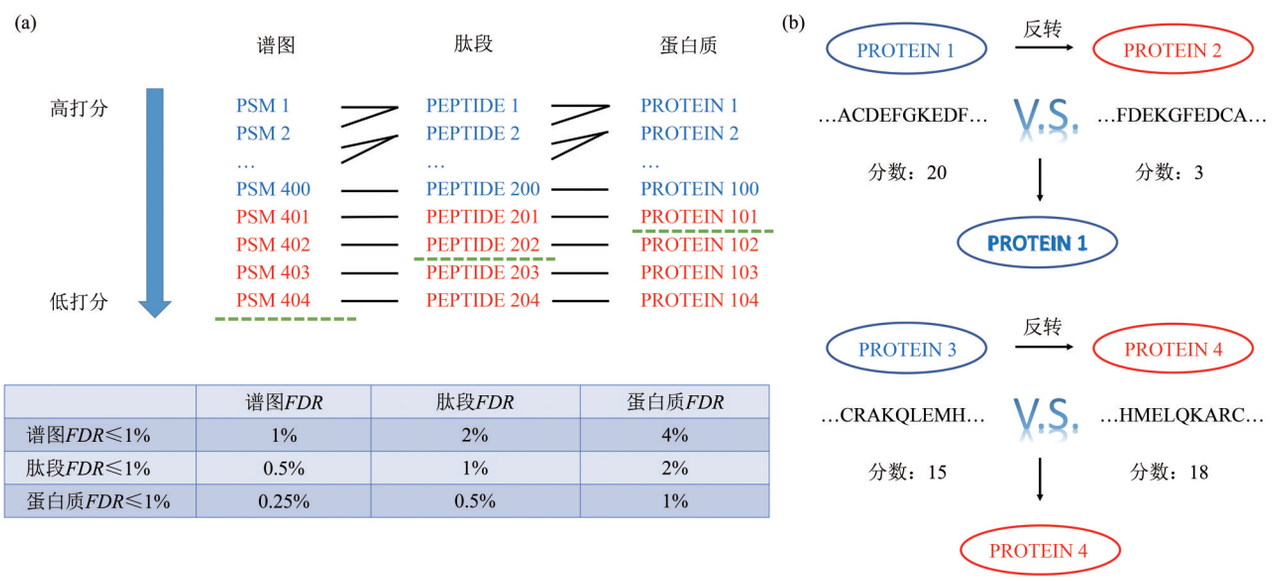
从而得到与公式(7)相同的计算公式[64]。所以，交联鉴定和糖肽鉴定中的TDA-FDR方法本质上是相同的。

2.2.3 针对蛋白质层面的TDA-FDR方法

质谱分析的终极目标是鉴定蛋白质。由谱图可以鉴定出肽段，进而推断出蛋白质，但这个向上递推的过程会导致错误结果逐渐积累[65-66]。例如，第一节提到的两篇人类蛋白质组草图研究中报告了多种错误蛋白质，其主要原因是这两篇草图文章都只对肽段的可信度进行了质控，没有对蛋白质层面做质量控制，在肽段推断蛋白质时，错误率得到了积累[26]。由于正确鉴定的肽段更有可能集中到相同的蛋白质，而错误鉴定的肽段则有可能分散到不同的蛋白质，这样就造成了从肽段推断到蛋白质后，蛋白质层面的错误率积累，造成蛋白质层面的*FDR*较高，是肽段层面的数倍或数十倍（图2a）。所以，从肽段推断到蛋白质后，还要对蛋白质层面进行质量控制。蛋白质的推断方式影响着蛋白质层面的质量控制，共享肽段的分配影响着蛋白质推断结果。有研究认为，蛋白质推断需要遵循奥卡姆剃刀原则，即用最少的蛋白质解释所有的肽段[67]。也有研究认为“one-hit-wonders”不可信[68-70]，需要引入双特异肽段推断方法，但Gupta等[71]认为双特异肽段推断过于保守。人类蛋白质组计划则明确表示鉴定遗漏蛋白质需要不低于9个氨基酸长度的非嵌套的双特异肽段[50-51]。

当蛋白质组数据集规模较大（能鉴定数十万条肽段）时，鉴定到的目标库蛋白质数目越来越多，造成新鉴定的目标库蛋白质和诱饵库蛋白质比例失衡，新鉴定的目标库蛋白质越来越少，新鉴定的诱饵库蛋白质越来越多，造成诱饵库蛋白质累积和蛋白质*FDR*的高估。针对大数据集带来的目标库和诱饵库蛋白质匹配概率失衡的问题，领域内目前发展了MAYU[72]和Picked FDR[73-74]等蛋白质推断及质控方法。这里介绍思想最简单、实现最方便又能取得较好效果的Picked FDR方法[73]，该方法将目标库蛋白质及其序列反转得到的诱饵库蛋白质看作一组，每组蛋白质中如果两个蛋白质都被鉴定，那么只保留打分高的蛋白质匹配，删除打分低的蛋白质匹配。在具体实现时可以将所有鉴定的蛋白质按照打分从高到低进行排序，对于每个蛋白质，如果其对应的反转蛋白质（目标库蛋白质的反转为诱饵蛋白质，诱饵蛋白质的反转为目标蛋白质）已经在前述蛋白质列表出现过，那么删除当前蛋白质，反之，则保留当前蛋白质。以图2b为例，目标库蛋白质PROTEIN 1获得了20分，其对应的诱饵库蛋白质PROTEIN 2获得了3分，那么打分高的PROTEIN 1被保留，打分低的PROTEIN 2被删除，不再参与后续蛋白质*FDR*计算。同理，目标蛋白质PROTEIN 3获得了15分，其对应的诱饵库蛋白质PROTEIN 4获得了18分，那么打分高的PROTEIN 4被保留，打分低的PROTEIN 3被删除。通过这种方法能够解决低打分区域鉴定到的目标库和诱饵库蛋白质数目不平衡的问题，使得TDA的1∶1假设在蛋白质层面得到满足，从而得到更准确的蛋白质*FDR*。Percolator 3.0文章中对Picked FDR方法进行了检验和肯定[75]。在Picked FDR原

理基础上，Prieto等[74]认为，诱饵库蛋白质的打分是无意义的，不应该删除比诱饵库蛋白质打分低的目标库蛋白质。所以，他们对Picked FDR方法做了改进，即对于打分低于目标库的诱饵库蛋白质予以删除，但对于打分低于诱饵库的目标库蛋白质予以保留。Prieto等认为改进的Picked FDR方法能够在保持与原Picked FDR方法相当的灵敏度的情况下，保留更多的高分蛋白质。



	谱图*FDR*	肽段*FDR*	蛋白质*FDR*
谱图*FDR*≤1%	1%	2%	4%
肽段*FDR*≤1%	0.5%	1%	2%
蛋白质*FDR*≤1%	0.25%	0.5%	1%

Fig. 2　Protein inference and protein level quality control

图2　蛋白质推断与质量控制

（a）从谱图推导肽段以及从肽段推断蛋白质都会导致错误率积累。图中用蓝色标注目标库鉴定结果，红色标注诱饵库鉴定结果，绿色虚线代表当前层面（谱图、肽段或者蛋白质）控制*FDR*为1%时的分界线。（b）Picked FDR方法。诱饵库蛋白质PROTEIN 2和PROTEIN 4分别由目标库蛋白质PROTEIN 1和PROTEIN 3的序列反转得到。

2.3　TDA-FDR方法的局限

TDA-FDR方法简单易用，并且能在子类肽段和交联肽段等特殊鉴定任务中演化出更合适的版本，但是该方法还存在两个局限。a. TDA-FDR方法估计的准确度有待考究。领域内普遍认为，目标库中的错误鉴定结果有两个来源：真正的随机匹配和同源错误匹配[24, 76]。当使用目标库序列反转或者随机化构建诱饵库序列时，TDA-FDR理论上能够模拟出随机匹配的分布情况，但却无法模拟出同源错误匹配情况，所以理论上TDA-FDR会低估真实的错误率[66]。另外，在二次搜索等特殊场景下，TDA-FDR会严重低估真实错误率，Jeong等[77]研究表明，在采用两步搜索方法对酵母数据进行搜索时，TDA方法估计的*FDR*是真实错误率的1/20。这可能是由于第二次搜索时采用第一次搜索鉴定的目标库蛋白质构造蛋白质小库，虽然通过目标库蛋白质序列反转构建了同等数目的诱饵库蛋白质，但此时的目标库蛋白质比诱饵库蛋白质更容易获得高分，造成TDA失衡。b. TDA-FDR方法不能对单个鉴定结果的可信度进行评价。Nesvizhskii[66]认为，TDA-FDR是全局方法，是对一组已经获取个体置信度分数的鉴定结果的假发现率进行的估计。鉴定结果的准确度会影响后续解析蛋白质结构与功能、研究致病机理和靶向治疗方案等工作的可行性和准确性。所以还需要在TDA-FDR方法的基础上，使用更严格的可信度评价方法，保证鉴定结果可以用于后续的结构和功能分析，这也是下一节提到的Beyond-TDA方法的由来。

3　Beyond-TDA方法

造成错误匹配的因素众多，搜索空间[12]、碎片离子强度[45, 76]和与实验参数相关的信息，如母离子误差、保留时间、酶切特异端点和遗漏酶切位点数目等[66]，都能帮助区分正确和错误鉴定结果。因此，在TDA方法的基础上，结合前述有效信息，可以进一步检验鉴定结果可信度。本文将这类方法

统称为Beyond-TDA方法，即在TDA-FDR方法的基础上，对鉴定结果的可信度做进一步检验。我们认为“评价”包含对群体鉴定可信度的评价（如TDA-FDR）和对个体鉴定可信度的评价，而本章介绍的Beyond-TDA方法均是对个体可信度的评价，即检验每个鉴定结果的正确性，所以我们又将其称为可信度检验方法。Beyond-TDA方法根据其使用的有效信息可以分为4类：a. 基于搜索空间的方法，包括陷阱库检验和开放式搜索检验；b. 基于谱图相似性的方法，包括合成肽段检验和理论谱图预测；c. 基于化学信息的方法，包括保留时间预测和同位素标记检验；d. 基于机器学习的方法，包括Percolator、pValid和DeepRescore等。

3.1 可信度检验方法的两个衡量指标

肽段鉴定可信度检验方法通常会给肽段的可信度进行打分，根据打分高低衡量不同肽段的可信程度。但是，在应用这些方法之前需要首先评估它们的检验能力，检验假阳率（false positive rate，*FPR*）和检验假阴率（false negative rate，*FNR*）就是这样两个衡量指标。在本文中，检验的目标就是为了发现错误鉴定结果，类似于临床中诊断疾病，患该疾病则为阳性，反之为阴性。所以本文将检验结果呈阳性定义为检验方法判断鉴定结果为错误鉴定，检验呈阴性则是指判断鉴定结果为正确，检验的假阳是指真实正确的鉴定结果被判断为阳性（错误鉴定），检验的假阴是指真实错误的鉴定结果被判断为阴性（正确鉴定）。进一步，检验假阳率是指正确鉴定结果被报告为不可信结果（即检验结果阳性）的比例，检验假阴率是指错误鉴定结果被报告为可信结果（即检验结果阴性）的比例[78]。从定义上看，这两个指标都是越小越好。同时，这些方法的检验假阳率和检验假阴率与应用它们排除检验阳性的结果前后鉴定结果的灵敏度和准确度存在一定的关系，即检验假阳率越低，排除检验阳性的结果后，鉴定结果的灵敏度越高，检验假阴率越低，排除检验阳性的结果后，鉴定结果的准确度越高[78]。检验假阳率和检验假阴率越低的方法对鉴定结果的正误判断越准确，在实际检验肽段鉴定可信度的过程中，应该选择检验假阳率和检验假阴率都较低的方法，保留检验方法认为可信的鉴定结果，排除它们认为不可信的鉴定结果。

3.2 基于搜索空间的可信度检验方法

搜索空间对鉴定结果的准确度有较大影响。当搜索空间不足即正确鉴定结果不在搜索空间内时，会导致鉴定出错。而扩大搜索空间，会有两种情况：第一，正确鉴定结果被包括到搜索空间中，只要肽段-谱图匹配打分无误，就可以鉴定到正确鉴定结果，将原始鉴定判错；第二，正确鉴定结果仍然不在搜索空间中，但此时搜索空间中更多的候选结果有更大的概率打败小空间搜索时的错误结果，这样也能评价原始鉴定结果的正确性。无论哪种情况，我们主要利用搜索空间增大后结果的不稳定性，对原始鉴定结果的可信度进行评价。根据搜索空间的不同扩增方式，又分为陷阱库检验和开放式搜索检验。

陷阱库方法已经在蛋白质组学研究中应用多年，其主要思想是使用与目标物种无关的蛋白质作为陷阱进行匹配，如果一张谱图在搜索目标蛋白质和陷阱蛋白质合并构成的数据库时匹配到陷阱库蛋白质的肽段，那么认为该谱图的鉴定结果是不可信的，这就可以用于评价不同引擎和不同方法的准确度[79-84]。马洁等[80]使用古细菌蛋白质库作为人类肝脏数据的陷阱库，比对搜索引擎在不同搜索参数下的错误率，提升搜索引擎的灵敏度和准确度。其实验结果表明，Mascot的Ion Score和Relative Score可以帮助提升鉴定灵敏度，使用贝叶斯非参数模型可以比根据人工经验确定的打分阈值过滤出的结果获得更高的准确度。Granholm等[81]使用流感嗜血杆菌蛋白质库作为18个ISB标准蛋白质的陷阱库，评测搜索引擎打分函数对正确和错误鉴定结果的区分能力。实验证明，使用了Intraset特征的Percolator的打分以及X！Tandem和MSGFDB中计算的*q*-value都是有偏的。Feng等[84]使用人类蛋白质库作为强烈火球菌的陷阱库蛋白质，使用古细菌蛋白质库作为人类数据的陷阱库，基于强烈火球菌和人类数据的陷阱库检验，评测了5种搜索引擎和四种质量控制方法，在这种评测条件下，搜索引擎MS-GF+和后处理方法PepDistiller[85]表现最优，同时也证明了使用分开过滤方法单独估计子类数据的*FDR*能够同时提升鉴定结果的准确度和灵敏度。具体使用陷阱库方法时有多种实现方式，选择不同物种、不同规模的蛋白质库作为陷阱库，会对实验结果造成不同程度的影响。Feng等[83]的研究指出，需要使用规模为目标蛋白质数据库十倍的陷阱库才能保证随机匹配几乎只发生在陷阱库上，使得陷阱库方法发挥最佳效果。

上述应用陷阱库思想的研究中都只搜索了目标库和陷阱库的合并蛋白质库，陷阱库可以帮助找出

搜索合并库时的一部分错误鉴定，但没法对常规情况下只搜索目标库时的鉴定结果做检验，所以我们前期的工作中提出了额外搜索合并库的陷阱库检验方法[78]。合并库中的陷阱库蛋白质扩大了搜索空间，如果搜索合并库时的鉴定结果与之前只搜索目标库时的鉴定结果不一致，则认为之前只搜索目标库时的鉴定结果错误。

开放式搜索检验与陷阱库检验的思想类似，都是通过扩大搜索空间后再次搜库，检验原始搜索空间鉴定的结果是否会产生变化。不同之处在于开放式搜索检验扩大的搜索空间中可能包含正确鉴定结果，但是实际操作中与陷阱库检验区别并不大。陷阱库检验需要额外搜索目标库和陷阱库的合并蛋白质库，而开放式搜索需要额外搜索目标物种库的所有酶切和所有修饰情况。由于开放式搜索空间包含真实正确鉴定结果，所以开放式搜索检验更容易发现原始结果中的错误，即开放式搜索检验方法的检验假阴率会优于陷阱库检验方法，我们前期的研究中也证明了这个结论[78]。

值得一提的是，前文提到的TDA方法，本质上也应用了扩大搜索空间的思想。实际上，如果首先仅搜索一次目标库，再搜索一次目标库和诱饵库的合并库，那么，TDA也是一种基于陷阱库的检验方法，诱饵库在这里起到陷阱库的作用，且目标库与陷阱库具有同规模的特点（也正是这一特点，可以在TDA方法基础上进行*FDR*估计）。具体讲，TDA用作检验方法时，假阳性结果是指将原本正确的鉴定结果检验为阳性即错误鉴定结果，也就是只搜索目标库时鉴定为目标库的正确结果，搜索目标库和诱饵库的合并库时鉴定为诱饵库结果。当然，这种可能性极小，原则上，如果真实结果存在于目标库中，那么一般认为诱饵库竞争不过目标库中的真实结果，因此可以认为TDA方法的检验假阳率是0。TDA方法检验所得的假阴性结果是指将原本错误的鉴定结果检验为阴性即正确鉴定结果，也就是搜索目标库时错误的目标库鉴定结果，在搜索目标库和诱饵库的合并库时，仍然鉴定到目标库结果，这种可能性是存在的，如果随机匹配到目标库和诱饵库的概率是1∶1的假设成立，那么可以认为TDA的检验假阴率是50%。

3.3　基于谱图相似性的可信度检验方法

搜索引擎对每个肽段-谱图匹配的打分其实就是对实验谱图与肽段的理论谱图的相似程度进行打分，理论谱图估计得越准确，打分的可信度越高。常规的数据库搜索引擎在生成肽段的理论谱图时，没有考虑碎片离子的强度信息，即给理论谱图中的所有碎片离子赋予相同的强度，这会造成一部分肽段-谱图匹配错误。在Beyond-TDA方法中，有一类方法通过在肽段-谱图匹配打分时考虑碎片离子强度，对鉴定的肽段的可信度进行评价，包括合成肽段检验和理论谱图预测两种方法。

3.3.1　合成肽段检验

合成肽段检验方法能够获取最真实和最精准的肽段理论谱图，所以合成肽段检验方法是领域内检验鉴定结果可信度的金标准。合成肽段检验方法常用来检验发现的新现象（比如新基因、遗漏注释蛋白质和新修饰），即对相应的新肽段进行合成，在尽可能相同的液相色谱条件和质谱仪参数等条件下打谱，通过计算新肽段对应的实验谱图与合成肽段对应的合成谱图的余弦相似度，判断新肽段的可信度[86-88]。一般以0.9作为合成肽段检验的余弦相似度阈值，达到或超过这个阈值则认为鉴定结果可信；反之，低于该阈值则认为鉴定结果不可信[87-88]。合成肽段检验方法是领域内目前公认的最好的个体可信度检验方法，我们前期的工作中评测合成肽段检验方法的*FPR*为0.06%，*FNR*为1.44%[78]。然而该方法的应用成本非常高，需要消耗时间和经济成本，难以大规模应用。

3.3.2　理论谱图预测

理论谱图预测方法可以看作是合成肽段检验的一种替代方法。采用机器学习特别是深度学习技术预测特定仪器、特定碎裂能量、特定电荷状态的肽段理论碎裂的谱图，将这类谱图称为预测谱图。常用的理论谱图预测软件有采用随机森林方法的MS2PIP[89]，采用双向长短期记忆网络的pDeep[90]、pDeep2[91]、DeepMass[92]和Guan_2019（关慎恒等人开发的软件）[93]以及采用双向递归循环神经网络的Prosit[94]等。与合成肽段检验方法类似，得到预测谱图后，计算实验谱图和预测谱图的余弦相似度，pValid文章中综合考虑*FPR*和*FNR*后选取0.7作为实验谱图和pDeep2预测谱图的相似度阈值，余弦相似度达到或者超过0.7认为鉴定结果可信，反之余弦相似度低于0.7则认为鉴定结果不可信，取阈值0.7时理论谱图预测方法的*FPR*和*FNR*分别是0.26%和10.80%[78]。理论谱图预测方法不仅可以用于检验鉴定结果的可信度，也可以帮助改进肽段和谱图的匹配打分，DeepMass从理论谱图中提取强度Top-3、Top-5、Top-7、Top-10和

Top-13的谱峰计算Andromeda打分[95]，虽然参与打分的谱峰数目比原始谱图要少，但是由于谱峰预测更准确，反而可以提升打分。

Xu等[96]对4种理论谱图预测软件的预测能力进行了评测，采用10个不同物种、酶切、仪器和碎裂能量的公共数据，对这些数据进行重新分析，采用5种数据库搜索引擎进行搜库，取搜库结果交集作为标注集，根据鉴定肽段的离子类型、长度和电荷进行分组，然后采用MS2PIP、Prosit、pDeep 2和Guan_2019预测肽段的理论谱图，计算理论谱图和实验谱图的皮尔逊相似度。从预测谱图与实验谱图的相似程度看，Prosit和pDeep2表现最好；从GPU和CPU上的运行时间看，pDeep2在GPU和CPU上运行时间均优于Prosit。

3.4 基于化学信息的可信度检验方法

除了上述基于搜索空间和谱图相似性的Beyond-TDA方法，引入保留时间和同位素标记等化学信息也可以帮助评价肽段的可信度。保留时间预测方法可以提供肽段的理论保留时间，而同位素标记方法相当于对待检验的目标增加了额外的谱图信息。

3.4.1 保留时间预测

肽段的保留时间是指肽段从色谱进入质谱所需要的时间，通俗来说是指肽段离子在质谱中从有信号到信号达到最高峰这段过程的时间，它与肽段的化学结构有关，在特定分离条件下肽段的保留时间应该是相对恒定的，所以通过检验肽段的保留时间是否在一定的范围内，就可以判断鉴定结果准确性[66, 97-98]。保留时间预测方法有采用支持向量回归方法的Elude[99-100]、采用高斯过程回归方法的GPTime[101]、采用胶囊网络和迁移学习方法的DeepRT[102]、采用双向递归循环神经网络的Prosit[94]、采用双向长短期记忆网络的Guan_2019[93]以及基于卷积神经网络和长短期记忆网络的AutoRT[103]。可以直接采用预测保留时间与实际保留时间的差值检验鉴定结果的可信度，也可以将差值作为一维特征，与理论谱图相似度等其他特征联合判断鉴定结果的可信度。

P-IVS[104]是一种结合合成肽段和保留时间特征的可信度检验方法，该方法对于感兴趣的目标肽段进行合成，同时在实验样品和合成样品中均混入一定量的标准肽段，统计标准肽段在两种样品中的谱图皮尔逊相似度和保留时间差值的范围，确定置信区间，然后对于目标肽段计算其在两种样品中的皮尔逊相似度和保留时间差值，通过前述确定的皮尔逊相似度和保留时间差值的置信区间对目标肽段的可信度进行检验。P-IVS的优势是结合了合成肽段和保留时间，使得可信度检验较为精准，但是不能大规模应用，文章中仅仅对11条目标肽段的可信度进行了检验（使用了40条标准肽段）。

3.4.2 同位素标记检验

同位素标记检验方法（图3）需要在样品制备过程中同时制备无标记样品和重同位素标记样品，将无标记样品和标记样品按比例混合后再进行酶切和质谱采集。对于搜索引擎鉴定的每条肽段，如果鉴定为无标记肽段，则在一级谱寻找其对应的重同位素标记肽段的信号峰；如果鉴定为重同位素标记肽段，则在一级谱寻找其对应的无标记肽段的信号峰。如果能找到该鉴定结果对应的另一种标记肽段的信号峰，那么认为该鉴定结果可信；反之，则认为该鉴定结果不可信[12, 61, 63]。更严格的同位素标记检验可以计算无标记和标记肽段信号峰的强度比值，只有比值符合或接近样品制备时无标记样品和标记样品的浓度比例，才认为鉴定结果可信，反之认为不可信。

同位素标记检验方法目前在常规蛋白质组学、交联蛋白质组学和糖蛋白质组学都得到了应用[12, 61, 63]。糖蛋白质组学中最早应用了同位素标记检验方法[63]，糖肽搜索引擎pGlyco 2首次应用^{15}N和^{13}C两种同位素标记方法标记酿酒酵母数据，其中无标记、^{15}N标记和^{13}C标记3种标记样品的比例是1∶1∶1。pGlyco 2和Byonic鉴定结果的^{15}N和^{13}C标记检验表明pGlyco 2的可信度远高于Byonic。pGlyco 2采用所有诱饵库结果估计同位素标记检验方法的检验假阴率，但没有估计检验假阳率，最后使用估计得到的检验假阴率对目标库结果的错误率做了校正，算得pGlyco 2鉴定到的糖肽的错误率低于Byonic。

在常规蛋白质组学中，开放式搜索引擎Open-pFind也应用了^{15}N和^{13}C两种同位素标记方法标记大肠杆菌数据[12]，其中无标记、^{15}N标记和^{13}C标记三种标记样品的比例是1∶1∶1。8种搜索引擎鉴定结果的^{15}N标记检验和^{13}C标记检验均表明Open-pFind的鉴定结果具有最高的准确度，且Open-pFind相对其他引擎单独鉴定的差集部分具有与交集部分相当的准确度。Open-pFind采用多引擎交集作为正样本估计两种同位素标记方法的检验假阳率，采用低打分区域的目标库鉴定结果作为负样

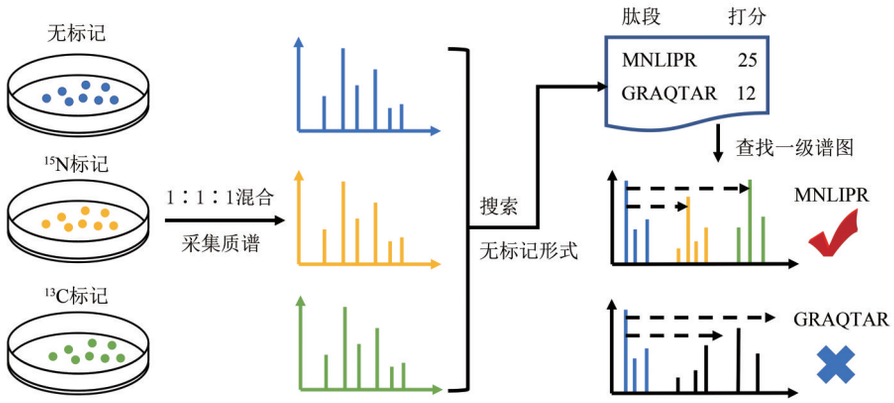


Fig. 3　Stable isotopic labeling validation method

图3　同位素标记检验方法

无标记和重同位素标记样品混合后采集质谱数据，对于搜索引擎给出的鉴定结果：如果鉴定为无标记肽段，则在一级谱寻找其对应的重同位素标记肽段的信号峰；如果鉴定为重同位素标记肽段，则在一级谱寻找其对应的无标记肽段的信号峰。如果能找到该鉴定结果对应的另一种标记肽段的信号峰，那么认为该鉴定结果可信；反之，则认为该鉴定结果不可信。

本估计两种同位素标记方法的检验假阴率，最后根据检验假阳率和检验假阴率估计出鉴定结果的错误率。

在交联蛋白质组学中，搜索引擎pLink 2的研究中采用^{15}N标记大肠杆菌数据[61]，无标记和^{15}N标记样品的比例是1∶1，分别采用两种交联剂Leiker和二硫键进行交联，在这两批交联剂数据上对三种交联引擎Kojak、pLink 1和pLink 2进行评测，检验结果表明pLink 2的鉴定结果具有最高的准确度。pLink 2采用多引擎交集作为正样本评测^{15}N标记检验方法的检验假阳率，采用通过TDA-FDR阈值的诱饵库鉴定结果作为负样本评测^{15}N标记检验方法的检验假阴率，最后根据检验假阳率和检验假阴率估计出pLink 2的鉴定结果在三个引擎的鉴定结果中错误率最低。

同位素标记检验方法的应用并不限于在常规蛋白质组学、交联蛋白质组学和糖蛋白质组学，还可以应用到微生物组学和蛋白质基因组学。同时，同位素标记检验方法也不限于MS1（一级质谱图）检验，还可以用于MS2（二级质谱图）检验，预期将有更高的检验效率。标记方法不限于^{15}N标记和^{13}C标记，其他代谢标记，如细胞培养条件下稳定同位素标记技术（stable isotope labeling by amino acids in cell culture，SILAC）和化学标记方法，都值得探索。

3.5　基于机器学习的可信度检验方法

上述3种基于搜索空间、谱图相似性和化学信息的Beyond-TDA方法都具有各自的优势，如果能结合以上3种方法的多种特征进行可信度检验，并结合机器学习等方法挖掘数据特性，将会得到更精准的可信度检验方法。Percolator[105]采用半监督学习方法，使得它能适配不同搜索引擎和不同物种的数据。Percolator采用互相关系数、质量、碎片离子匹配率、酶切特异性、肽段长度、电荷和鉴定结果数目等未用于打分的特征，使用支持向量机（support vector machine，SVM）作为分类器，对鉴定结果进行重打分。重打分的目的是为了让目标库和诱饵库结果区分度更高，达到检验鉴定结果可信度的目的。

DeepRescore[106]使用AutoRT[103]预测保留时间，计算预测保留时间与实验保留时间的差值DeltaRT，同时使用pDeep2预测理论谱图，计算理论谱图和实验谱图之间的谱图夹角（spectra angle，SA），将DeltaRT和SA作为特征加入Percolator，同搜索引擎给出的打分等特征一起重新训练，对每个鉴定结果重新打分，并重新计算*FDR*。

pValid方法从开放式搜索及理论谱图预测中提取与鉴定结果相关的特征，并采用SVM方法作为分类器，对鉴定结果的可信度进行预测[78]。开放式搜索同时考虑了特异、半特异、非特异酶切形式以及Unimod[107]中的所有修饰，也是一种扩大搜

索空间的检验方法。pValid综合了开放式搜索和理论谱图预测两种可信度检验方法，获得了更低的检验假阳率和检验假阴率，我们前期的工作中对以上提到的陷阱库、开放式搜索、合成肽段、理论谱图预测和pValid方法的检验假阳率和检验假阴率进行了研究[78]。采用3种数据库搜索引擎（pFind、MaxQuant和PEAKS）的交集构建正样本，评测各种方法的检验假阳率，采用正样本谱图母离子偏离5 u和10 u构建诱饵谱图重新搜库的方法构建负样本，评测各种方法的检验假阴率。在3批标注数据集上，pValid的检验假阴率最低，检验假阳率仅次于陷阱库方法。pValid的平均检验假阳率为0.03%，陷阱库方法的平均检验假阳率为0.01%，pValid的平均检验假阴率为1.79%，但陷阱库方法的平均检验假阴率高达56.13%。综合考虑检验假阳率和检验假阴率，pValid方法优于陷阱库、开放式搜索和理论谱图预测方法。在合成肽段数据集上，pValid的检验假阳率和检验假阴率媲美合成肽段检验方法（表1）。可以认为基于机器学习的pValid方法在一定条件下超越了陷阱库、开放式搜索和理论谱图预测方法，甚至也超越了合成肽段检验方法。

Table 1　Beyond-TDA validation methods

表1　Beyond-TDA方法

Beyond-TDA方法		核心思想	检验假阳率/%	检验假阴率/%
搜索空间	陷阱库检验	搜索目标物种和无关物种的合并蛋白质数据库，如果与仅搜索目标物种蛋白质库的结果不同，则认为仅搜索目标物种蛋白质库时的结果错误	0.01#	56.13#
	开放式搜索检验	搜索目标物种的所有酶切和所有修饰形式，其余步骤与陷阱库检验类似	0.04#	49.05#
谱图相似性	合成肽段检验	合成鉴定到的肽段，采集合成谱图，与实验谱图计算相似度	0.06*	1.44*
	理论谱图预测	合成肽段替代方法，采用深度学习技术预测肽段的理论谱图并与实验谱图计算相似度	0.26#	10.80#
化学信息	保留时间预测	预测肽段的保留时间，与实际保留时间做差，可与其他方法联用	/	/
	同位素标记检验	将无标记及同位素标记样品混合后酶切及采集质谱，对鉴定结果寻找一级谱图无标记、重同位素标记峰簇，进一步可以计算比值	/	/
机器学习	Percolator	采用互相关系数、质量、碎片离子匹配率、酶切特异性、肽段长度、电荷和鉴定结果数目等未用于打分的特征，使用SVM做半监督学习	/	/
	DeepRescore	将保留时间预测和理论谱图预测的两个特征加入Percolator做重打分	/	/
	pValid	采用开放式搜索检验和理论谱图预测相关的四个特征，使用SVM重打分	0.03# 0.04*	1.79# 1.44*

#表示在三批标注集数据上的平均检验假阳率和检验假阴率，*表示仅在合成肽段标注数据集上的检验假阳率和检验假阴率，/表示领域内暂无关于该方法*FDR*和*FNR*的定量分析。

4　总结与展望

质谱分析对蛋白质组学至关重要。质谱数据鉴定结果能够给出基因表达的直接证据，同时帮助解析蛋白质的结构和功能，发现与疾病相关的基因和蛋白质以及研制靶向治疗方案。然而，质谱分析结果的可信度亟待评价。对常规肽段使用TDA进行质量控制的方法虽然在子类肽段和交联肽段中都进行了演化改进，但仍然存在估计值不准确以及无法评价单个鉴定结果可信度的局限。因此，领域内在TDA基础上开发了结合搜索空间、谱图相似性、化学信息和机器学习等有效手段的Beyond-TDA方法。

Beyond-TDA方法主要介绍了基于搜索空间、谱图相似性和化学信息的3类方法，包括陷阱库、开放式搜索、合成肽段、理论谱图预测、保留时间预测和同位素标记检验方法。陷阱库方法可以快速检验大规模鉴定结果，TDA方法本质上也可以看作是陷阱库检验。开放式搜索也是一种扩大搜索空间的检验方法，因其扩大的空间中可能包含正确鉴定结果，所以它的检验假阴率理论上会优于陷阱库方法。合成肽段方法是检验金标准，但是时间和经济成本较高，不适用于大规模质谱数据鉴定结果的检验，由此产生了理论谱图预测方法模拟和替代合成肽段方法。保留时间预测方法采用预测保留时间与实际保留时间的差值作为鉴定结果可信度的评判标准，常常与理论谱图预测等方法联用。同位素标记检验目前已经在常规蛋白质组学、交联蛋白质组

学和糖蛋白质组学中得到了应用并发挥了重要价值，但这种方法还可以继续改进，比如不仅仅考虑无标记和重同位素标记肽段信号峰的存在性，将肽段的同位素峰簇比值以及碎片离子同位素峰簇比值都纳入检验范围，以及进一步从MS1拓展到MS2，从^{15}N和^{13}C拓展到SILAC，从代谢标记拓展到化学标记。

基于机器学习的可信度评价方法主要用于对鉴定结果进行重打分，自动选择最优重打分阈值检验鉴定结果的可信度，这些方法各自使用了多种特征，比如Percolator使用了XCorr互相关系数、肽段长度、电荷、鉴定结果数目等肽段-谱图匹配相关的特征，DeepRescore使用了保留时间差值和理论谱图预测，pValid使用了开放式搜索和理论谱图预测。这些特征都能帮助区分正确和错误鉴定结果，未来可以将这些特征综合应用到一个分类器中，并结合深度学习带来的优势，提升分类结果的准确性。未来也可以考虑结合所有Beyond-TDA方法的优势，构建更准确的可信度评价方法。需要注意的是，机器学习方法受限于训练数据的规模和质量，产生质谱数据的真实肽段是未知的，可以通过取多种搜索引擎交集的方法构建大规模高质量的正确鉴定结果，但构建同样规模的高质量的错误鉴定结果却很困难，这也是未来需要解决的问题。我们也注意到，近年来有研究认为，随着质谱仪精度越来越高，基于统计的方法（*P*-value和Benjamini-Hochberg方法[108]）的准确度优于常规的TDA-FDR[109]，理论上，这类方法也可以和文中提到的其他Beyond-TDA方法进行结合，进一步检验鉴定结果的可信度。

蛋白质组学领域内发展了TDA方法和基于搜索空间、谱图相似性、化学信息和机器学习技术的Beyond-TDA方法，对肽段鉴定的可信度进行评价，但是对于蛋白质层面的可信度评价关注不算多。蛋白质作为质谱分析的最终目标，具有非常重要的意义。Picked FDR方法让人们意识到日益增长的蛋白质组学数据中的蛋白质*FDR*高估问题，给出了简便且有效的解决方法，未来还需要更多地关注蛋白质层面的可信度评价方法。未来也可以尝试将目前的肽段可信度评价方法迁移和拓展到蛋白质的可信度评价中，比如，对于每个待评价的蛋白质，只要有一条特异的肽段通过了可信度评价，那么就可以认为此蛋白质也通过了可信度评价，具体实现方式与方法可行性还有待进一步分析探索。

参　考　文　献

[1] Aebersold R, Mann M. Mass spectrometry-based proteomics. Nature, 2003, **422**(6928): 198-207

[2] Kim M S, Pinto S M, Getnet D, *et al*. A draft map of the human proteome. Nature, 2014, **509**(7502): 575-581

[3] Wilhelm M, Schlegl J, Hahne H, *et al*. Mass-spectrometry-based draft of the human proteome. Nature, 2014, **509**(7502): 582-587

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

[5] Singh P, Panchaud A, Goodlett D R. Chemical cross-linking and mass spectrometry as a low-resolution protein structure determination technique. Ana Chem, 2010, **82**(7): 2636-2642

[6] Yang B, Wu Y J, Zhu M, *et al*. Identification of cross-linked peptides from complex samples. Nat Methods, 2012, **9**(9): 904-906

[7] Sun W, Xing B C, Sun Y, *et al*. Proteome analysis of hepatocellular carcinoma by two-dimensional difference gel electrophoresis. Mol Cell Proteomics, 2007, **6**(10): 1798-1808

[8] Ge S, Xia X, Ding C, *et al*. A proteomic landscape of diffuse-type gastric cancer. Nat Commun, 2018, **9**(1): 1012

[9] Jiang Y, Sun A H, Zhao Y, *et al*. Proteomics identifies new therapeutic targets of early-stage hepatocellular carcinoma. Nature, 2019, **567**(7747): 257-261

[10] Wang L H, Li D Q, Fu Y, *et al*. pFind 2.0: a software package for peptide and protein identification *via* tandem mass spectrometry. Rapid Commun Mass Spectrom, 2007, **21**(18): 2985-2991

[11] Cox J, Mann M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat Biotechnol, 2008, **26**(12): 1367-1372

[12] Chi H, Liu C, Yang H, *et al*. Comprehensive identification of peptides in tandem mass spectra using an efficient open search engine. Nat Biotechnol, 2018, **36**(11): 1059-1061

[13] Ma B, Zhang K Z, Hendrie C, *et al*. PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. Rapid Commun Mass Spectrom, 2003, **17**(20): 2337-2342

[14] Zhang J, Xin L, Shan B Z, *et al*. PEAKS DB: *de novo* sequencing assisted database search for sensitive and accurate peptide identification. Mol Cell Proteomics, 2012, **11**(4): M111.010587

[15] Chi H, Chen H F, He K, *et al*. pNovo+: *de novo* peptide sequencing using complementary HCD and ETD tandem mass spectra. J Proteome Res, 2013, **12**(2): 615-625

[16] Yang H, Chi H, Zhou W J, *et al*. Open-pNovo: *de novo* peptide sequencing with thousands of protein modifications. J Proteome Res, 2017, **16**(2): 645-654

[17] 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

[18] 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-i406

[19] 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

[20] Shen C, Wang Z, Shankar G, *et al*. A hierarchical statistical model to assess the confidence of peptides and proteins inferred from tandem mass spectrometry. Bioinformatics, 2008, **24**(2): 202-208

[21] Shi J, Chen B, Wu F X. Unifying protein inference and peptide identification with feedback to update consistency between peptides. Proteomics, 2013, **13**(2): 239-247

[22] The M, Edfors F, Perez-Riverol Y, *et al*. A protein standard that emulates homology for the characterization of protein inference algorithms. J Proteome Res, 2018, **17**(5): 1879-1886

[23] Nesvizhskii A I, Roos F F, Grossmann J, *et al*. Dynamic spectrum quality assessment and iterative computational analysis of shotgun proteomic data - toward more efficient identification of post-translational modifications, sequence polymorphisms, and novel peptides. Mol Cell Proteomics, 2006, **5**(4): 652-670

[24] Chen Y, Zhang J M, Xing G, *et al*. Mascot-derived false positive peptide identifications revealed by manual analysis of tandem mass spectra. J Proteome Res, 2009, **8**(6): 3141-3147

[25] Ezkurdia I, Vazquez J, Valencia A, *et al*. Analyzing the first drafts of the human proteome. J Proteome Res, 2014, **13**(8): 3854-3855

[26] Ezkurdia I, Calvo E, Del Pozo A, *et al*. The potential clinical impact of the release of two drafts of the human proteome. Expert Rev Proteomics, 2015, **12**(6): 579-593

[27] Garbers C, Rose-John S. Pharmaceutical relevant cytokine receptors: lessons from the first drafts of the human proteome. J Proteome Res, 2015, **14**(2): 1330-1332

[28] Verbeurgt C, Wilkin F, Tarabichi M, *et al*. Profiling of olfactory receptor gene expression in whole human olfactory mucosa. PLoS One, 2014, **9**(5): e96333

[29] Resing K A, Meyer-Arendt K, Mendoza A M, *et al*. Improving reproducibility and sensitivity in identifying human proteins by shotgun proteomics. Anal Chem, 2004, **76**(13): 3556-3568

[30] Corbett B A, Kantor A B, Schulman H, *et al*. A proteomic study of serum from children with autism showing differential expression of apolipoproteins and complement proteins. Mol Psychiatry, 2007, **12**(3): 292-306

[31] Fenyo D, Beavis R C. A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. Anal Chem, 2003, **75**(4): 768-774

[32] Keller A, Nesvizhskii A I, Kolker E, *et al*. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. Anal Chem, 2002, **74**(20): 5383-5392

[33] Nesvizhskii A I, Keller A, Kolker E, *et al*. A statistical model for identifying proteins by tandem mass spectrometry. Anal Chem, 2003, **75**(17): 4646-4658

[34] Choi H, Nesvizhskii A I. Semisupervised model-based validation of peptide identifications in mass spectrometry-based proteomics. J Proteome Res, 2008, **7**(1): 254-265

[35] Choi H, Ghosh D, Nesvizhskii A I. Statistical validation of peptide identifications in large-scale proteomics using the target-decoy database search strategy and flexible mixture modeling. J Proteome Res, 2008, **7**(1): 286-292

[36] Ding Y, Choi H, Nesvizhskii A I. Adaptive discriminant function analysis and reranking of MS/MS database search results for improved peptide identification in shotgun proteomics. J Proteome Res, 2008, **7**(11): 4878-4889

[37] Shteynberg D, Deutsch E W, Lam H, *et al*. iProphet: multi-level integrative analysis of shotgun proteomic data improves peptide and protein identification rates and error estimates. Mol Cell Proteomics, 2011, **10**(12): M111.007690

[38] Moore R E, Young M K, Lee T D. Qscore: an algorithm for evaluating SEQUEST database search results. J Am Soc Mass Spectrom, 2002, **13**(4): 378-386

[39] Peng J M, Elias J E, Thoreen C C, *et al*. Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome. J Proteome Res, 2003, **2**(1): 43-50

[40] Huttlin E L, Hegeman A D, Harms A C, *et al*. Prediction of error associated with false-positive rate determination for peptide identification in large-scale proteomics experiments using a combined reverse and forward peptide sequence database strategy. J Proteome Res, 2007, **6**(1): 392-398

[41] Haas W, Faherty B K, Gerber S A, *et al*. Optimization and use of peptide mass measurement accuracy in shotgun proteomics. Mol Cell Proteomics, 2006, **5**(7): 1326-1337

[42] Bianco L, Mead J A, Bessant C. Comparison of novel decoy database designs for optimizing protein identification searches using ABRF sPRG2006 standard MS/MS data sets. J Proteome Res, 2009, **8**(4): 1782-1791

[43] Wang G, Wu W W, Zhang Z, *et al*. Decoy methods for assessing false positives and false discovery rates in shotgun proteomics. Anal Chem, 2009, **81**(1): 146-159

[44] Keich U, Noble W S. Progressive calibration and averaging for tandem mass spectrometry statistical confidence estimation: why settle for a single decoy?. Res Comput Mol Biol, 2017, **10229**: 99-116

[45] Keich U, Noble W S. Controlling the FDR in imperfect matches to an incomplete database. J Am Stat Assoc, 2018, **113**(523): 973-982

[46] Keich U, Tamura K, Noble W S. Averaging strategy to reduce variability in target-decoy estimates of false discovery rate. J Proteome Res, 2019, **18**(2): 585-593

[47] Kall L, Storey J D, Noble W S. Non-parametric estimation of posterior error probabilities associated with peptides identified by tandem mass spectrometry. Bioinformatics, 2008, **24**(16): i42-i48

[48] 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

[49] Jiang X N, Dong X L, Ye M L, *et al*. Instance based algorithm for posterior probability calculation by target-decoy strategy to improve protein identifications. Anal Chem, 2008, **80**(23): 9326-9335

[50] Deutsch E W, Overall C M, Van Eyk J E, *et al*. Human proteome project mass spectrometry data interpretation guidelines 2.1. J

Proteome Res, 2016, **15**(11): 3961-3970

[51] Deutsch E W, Lane L, Overall C M, *et al*. Human proteome project mass spectrometry data interpretation guidelines 3.0. J Proteome Res, 2019, **18**(12): 4108-4116

[52] Fu Y. Bayesian false discovery rates for post-translational modification proteomics. Stat Interface, 2012, **5**(1): 47-59

[53] Branca R M M, Orre L M, Johansson H J, *et al*. HiRIEF LC-MSMS enables deep proteome coverage and unbiased proteogenomics. Nat Methods, 2014, **11**(1): 59-62

[54] Zhang K, Fu Y, Zeng W, *et al*. A note on the false discovery rate of novel peptides in proteogenomics. Bioinformatics, 2015, **31**(20): 3249-3253

[55] Li J, Su Z, Ma Z, *et al*. A bioinformatics workflow for variant peptide detection in shotgun proteomics. Mol Cell Proteomics, 2011, **10**(5): M110.006536

[56] Fu Y, Qian X H. Transferred subgroup false discovery rate for rare post-translational modifications detected by mass spectrometry. Mol Cell Proteomics, 2014, **13**(5): 1359-1368

[57] Tariq M U, Haseeb M, Aledhari M, *et al*. Methods for proteogenomics data analysis, challenges, and scalability bottlenecks: a survey. Ieee Access, 2021, **9**: 5497-5516

[58] Krug K, Carpy A, Behrends G, *et al*. Deep coverage of the *Escherichia coli* proteome enables the assessment of false discovery rates in simple proteogenomic experiments. Mol Cell Proteomics, 2013, **12**(11): 3420-3430

[59] 张昆 . 蛋白质基因组学新基因发现与验证策略研究[D]. 北京: 中国科学院大学, 2015
Zhang K. Discovery and Validation of Novel Genes in Proteogenomics[D]. Beijing: University of Chinese Academy of Sciences, 2015

[60] Walzthoeni T, Claassen M, Leitner A, *et al*. False discovery rate estimation for cross-linked peptides identified by mass spectrometry. Nat Methods, 2012, **9**(9): 901-903

[61] Chen Z L, Meng J M, Cao Y, *et al*. A high-speed search engine pLink 2 with systematic evaluation for proteome-scale identification of cross-linked peptides. Nat Commun, 2019, **10**: 3404

[62] Strum J S, Nwosu C C, Hua S, *et al*. Automated assignments of N- and O-site specific glycosylation with extensive glycan heterogeneity of glycoprotein mixtures. Anal Chem, 2013, **85**(12): 5666-5675

[63] Liu M Q, Zeng W F, Fang P, *et al*. pGlyco 2.0 enables precision N-glycoproteomics with comprehensive quality control and one-step mass spectrometry for intact glycopeptide identification. Nat Commun, 2017, **8**: 438

[64] 曾文锋 . 基于生物质谱技术的规模化完整糖肽鉴定方法研究[D]. 北京: 中国科学院大学, 2016
Zeng W F. Mass Spectrometry-based Large-Scale Identification of Intact Glycopeptides[D]. Beijing: University of Chinese Academy of Sciences, 2016

[65] 李宁, 吴松锋, 朱云平, 等 . 鸟枪法蛋白质鉴定质量控制方法研究进展 . 生物化学与生物物理进展, 2009, **36**(6): 668-675
Li N, Wu S F, Zhu Y P, *et al*. Prog Biochem Biophys, 2009, **36**(6): 668-675

[66] Nesvizhskii A I. A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. J Proteomics, 2010, **73**(11): 2092-2123

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

[68] Carr S, Aebersold R, Baldwin M, *et al*. The need for guidelines in publication of peptide and protein identification data - working group on publication guidelines for peptide and protein identification data. Mol Cell Proteomics, 2004, **3**(6): 531-533

[69] Omenn G S, States D J, Adamski M, *et al*. Overview of the HUPO Plasma Proteome Project: results from the pilot phase with 35 collaborating laboratories and multiple analytical groups, generating a core dataset of 3020 proteins and a publicly-available database. Proteomics, 2005, **5**(13): 3226-3245

[70] Bradshaw R A, Burlingame A L, Carr S, *et al*. Reporting protein identification data - the next generation of guidelines. Mol Cell Proteomics, 2006, **5**(5): 787-788

[71] Gupta N, Pevzner P A. False discovery rates of protein identifications: a strike against the two-peptide rule. J Proteome Res, 2009, **8**(9): 4173-4181

[72] Reiter L, Claassen M, Schrimpf S P, *et al*. Protein identification false discovery rates for very large proteomics data sets generated by tandem mass spectrometry. Mol Cell Proteomics, 2009, **8**(11): 2405-2417

[73] Savitski M M, Wilhelm M, Hahne H, *et al*. A scalable approach for protein false discovery rate estimation in large proteomic data sets. Mol Cell Proteomics, 2015, **14**(9): 2394-2404

[74] Prieto G, Vazquez J. Protein probability model for high-throughput protein identification by mass spectrometry-based proteomics. J Proteome Res, 2020, **19**(3): 1285-1297

[75] The M, Maccoss M J, Noble W S, *et al*. Fast and accurate protein false discovery rates on large-scale proteomics data sets with percolator 3.0. J Am Soc Mass Spectrom, 2016, **27**(11): 1719-1727

[76] Choi H, Nesvizhskii A I. False discovery rates and related statistical concepts in mass spectrometry-based proteomics. J Proteome Res, 2008, **7**(1): 47-50

[77] Jeong K, Kim S, Bandeira N. False discovery rates in spectral identification. BMC Bioinformatics, 2012, **13** (Suppl 16): S2

[78] Zhou W J, Yang H, Zeng W F, *et al*. pValid: validation beyond the target-decoy approach for peptide identification in shotgun proteomics. J Proteome Res, 2019, **18**(7): 2747-2758

[79] Zhang J Y, Ma J, Dou L, *et al*. Bayesian nonparametric model for the validation of peptide identification in shotgun proteomics. Mol Cell Proteomics, 2009, **8**(3): 547-557

[80] Ma J, Zhang J Y, Wu S F, *et al*. Improving the sensitivity of MASCOT search results validation by combining new features with Bayesian nonparametric model. Proteomics, 2010, **10**(23): 4293-4300

[81] Granholm V, Noble W S, Kall L. On using samples of known protein content to assess the statistical calibration of scores assigned to peptide-spectrum matches in shotgun proteomics. J

Proteome Res, 2011, **10**(5): 2671-2678

[82] Vaudel M, Burkhart J M, Breiter D, *et al*. A complex standard for protein identification, designed by evolution. J Proteome Res, 2012, **11**(10): 5065-5071

[83] Feng X D, Ma J, Chang C, *et al*. An improved target-decoy strategy for evaluation of database search engines and quality control methods in shotgun proteomics//Ching C T S. Proceedings of the 2016 International Conference on Biomedical and Biological Engineering. Paris: Atlantis Press, 2016: 366-372

[84] Feng X D, Li L W, Zhang J H, *et al*. Using the entrapment sequence method as a standard to evaluate key steps of proteomics data analysis process. BMC Genomics, 2017, **18**(2): 143

[85] Li N, Wu S F, Zhang C P, *et al*. PepDistiller: a quality control tool to improve the sensitivity and accuracy of peptide identifications in shotgun proteomics. Proteomics, 2012, **12**(11): 1720-1725

[86] Quandt A, Espona L, Balasko A, *et al*. Using synthetic peptides to benchmark peptide identification software and search parameters for MS/MS data analysis. Eupa Open Proteomics, 2014, **5**: 21-31

[87] Peng X H, Xu F, Liu S, *et al*. Identification of missing proteins in the phosphoproteome of kidney cancer. J Proteome Res, 2017, **16**(12): 4364-4373

[88] Wang Y H, Chen Y, Zhang Y, *et al*. Multi-protease strategy identifies three PE2 missing proteins in human testis tissue. J Proteome Res, 2017, **16**(12): 4352-4363

[89] Degroeve S, Martens L. (MSPIP) -P-2: a tool for MS/MS peak intensity prediction. Bioinformatics, 2013, **29**(24): 3199-3203

[90] 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

[91] Zeng W F, Zhou X X, Zhou W J, *et al*. MS/MS spectrum prediction for modified peptides using pDeep2 trained by transfer learning. Anal Chem, 2019, **91**(15): 9724-9731

[92] 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

[93] Guan S H, Moran M F, Ma B. Prediction of LC-MS/MS properties of peptides from sequence by deep learning. Mol Cell Proteomics, 2019, **18**(10): 2099-2107

[94] 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

[95] Cox J, Neuhauser N, Michalski A, *et al*. Andromeda: a peptide search engine integrated into the MaxQuant environment. J Proteome Res, 2011, **10**(4): 1794-1805

[96] Xu R, Sheng J, Bai M Z, *et al*. A comprehensive evaluation of MS/MS spectrum prediction tools for shotgun proteomics. Proteomics, 2020, **20**(21-22): 1900345

[97] Strittmatter E F, Kangas L J, Petritis K, *et al*. Application of peptide LC retention time information in a discriminant function for peptide identification by tandem mass spectrometry. J Proteome Res, 2004, **3**(4): 760-769

[98] Baczek T, Kaliszan R. Predictions of peptides' retention times in reversed-phase liquid chromatography as a new supportive tool to improve protein identification in proteomics. Proteomics, 2009, **9**(4): 835-847

[99] Moruz L, Staes A, Foster J M, *et al*. Chromatographic retention time prediction for posttranslationally modified peptides. Proteomics, 2012, **12**(8): 1151-1159

[100] Moruz L, Tomazela D, Kall L. Training, selection, and robust calibration of retention time models for targeted proteomics. J Proteome Res, 2010, **9**(10): 5209-5216

[101] Afkham H M, Qiu X B, The M, *et al*. Uncertainty estimation of predictions of peptides' chromatographic retention times in shotgun proteomics. Bioinformatics, 2017, **33**(4): 508-513

[102] Ma C W, Ren Y, Yang J R, *et al*. Improved peptide retention time prediction in liquid chromatography through deep learning. Anal Chem, 2018, **90**(18): 10881-10888

[103] Wen B, Li K, Zhang Y, *et al*. Cancer neoantigen prioritization through sensitive and reliable proteogenomics analysis. Nat Commun, 2020, **11**(1): 1759

[104] Wiles T A, Saba L M, Delong T. Peptide-spectrum match validation with internal standards (p-vis): internally-controlled validation of mass spectrometry-based peptide identifications. J Proteome Res, 2021, **20**(1): 236-249

[105] 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

[106] Li K, Jain A, Malovannaya A, *et al*. DeepRescore: leveraging deep learning to improve peptide identification in immunopeptidomics. Proteomics, 2020, **20**(21-22): 1900334

[107] Creasy D M, Cottrell J S. Unimod: protein modifications for mass spectrometry. Proteomics, 2004, **4**(6): 1534-1536

[108] Benjamini Y, Hochberg Y. Controlling the false discovery rate - a practical and powerful approach to multiple testing. J R Stat Soc B, 1995, **57**(1): 289-300

[109] Coute Y, Bruley C, Burger T. Beyond target-decoy competition: stable validation of peptide and protein identifications in mass spectrometry-based discovery proteomics. Anal Chem, 2020, **92**(22): 14898-14906

Validation Methods of Peptide Identification Results in Proteomics*

ZHOU Wen-Jing[1,2], ZENG Wen-Feng[1], CHI Hao[1,2]**, HE Si-Min[1,2]**

([1]*Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences, Institute of Computing Technology, Chinese Academy of Sciences, Beijing* 100190, China;

[2]*University of Chinese Academy of Sciences, Beijing* 100049, China)

Abstract　Mass spectrometry-based proteomics aims to identify peptides and proteins to give direct proofs of gene expressions, analyze structures and functions of proteins, study the relationship between proteins and diseases, and provide targeted treatment options. All these studies are based on the credibility of identified peptides and proteins. However, it is impossible to manually check all identified peptides because a large number of identifications can be collected from one mass spectrometry experiment. Thus, target-decoy approach (TDA) is proposed and always used to control the quality of identified peptides and proteins, and has been expanded to subclasses of peptides (including ordinary subclasses of peptides, variant peptides, and modified peptides) and cross-linking peptides. However, TDA still has two limitations: (1) the estimation of false discovery rate (*FDR*) is inaccurate and (2) validation of single identification cannot be supported. Thus, the identification results that passed the TDA-based FDR control need to be further validated and other validation methods which are used after TDA-FDR filtration (referred to as Beyond-TDA methods) have been developed to enhance peptide validation. This paper reviews TDA and its extensions as well as Beyond-TDA methods and discusses the advantages and disadvantages of each method. In the first part of this paper, we introduce the goal of proteomics, the process of mass spectrometry acquisition and analysis, the validation problem, and the early statistical methods to evaluate the identification credibility. Then, in the second part of this paper, we describe in detail the ordinary TDA-FDR method, including the assumption that random matches are equally likely to appear in target and decoy databases, the construction methods to generate the decoy database, and the computational formula of TDA-FDR. We also introduce the extensions of TDA-FDR on ordinary subclasses of peptides, variant peptides, modified peptides, proteogenomics peptides, cross-linking peptides, and glycopeptides. However, TDA cannot model the homologous incorrect peptides, thus TDA-FDR underestimates the actual false rate. So, after TDA-FDR filtration, it is necessary to use more strict validation methods, *i.e.*, Beyond-TDA methods, which are reviewed in detail in the third part of this paper, to control validation credibility. In this part, four kinds of methods are introduced, including validation methods based on search space (trap database validation and open search validation), spectra similarity (synthetic peptide validation and theoretical spectra prediction), chemical information (retention time prediction and stable isotopic labeling validation) and machine learning technology (Percolator, pValid, and DeepRescore). Lastly, we summarize the content of this paper and discuss the future improvement directions of validation methods.

Key words　proteomics, mass spectrometry, target-decoy approach, false discovery rate, validation methods

DOI：10.16476/j.pibb.2022.0004

* This work was supported by grants from the National Key Research and Development Program of China (2016YFA0501300) and The National Nature Science Foundation of China Excellent Young Scientists Fund Program (32022046).

** Corresponding author.

CHI Hao. Tel：86-10-62600822，E-mail：chihao@ict.ac.cn

HE Si-Min. Tel：86-10-62600822，E-mail：smhe@ict.ac.cn

Received：January 6，2022　Accepted：March 23，2022