PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
www.pibb.ac.cn

肽段的理论串联质谱图预测方法研究进展*

周撷璇[1,3] 任 睿[1,3] 高婉铃[1,3] 黄运有[1,3] 曾文锋[2,3]
孔德飞[1,3] 郝天舒[1,3] 张知非[4)**] 詹剑锋[1,3)**]

([1)] 中国科学院计算技术研究所，中国科学院计算机体系结构国家重点实验室，北京 100190；
[2)] 中国科学院计算技术研究所，中国科学院智能信息处理重点实验室，北京 100190；
[3)] 中国科学院大学，北京 100049；[4)] 首都医科大学，北京 200031）

摘要 基于串联质谱技术的蛋白质组学已经成为生命科学领域的重要工具，其中肽段的理论串联质谱图（通常也被称为二级谱图）预测问题在近年来广受关注. 大量高质量质谱数据的积累和计算技术的发展为此问题的解决提供了有效途径. 肽段的理论二级谱图预测的方法可以分为两大类，一类是基于物理模型的方法，即基于移动质子模型的方法，例如MassAnalyzer、MS-Simulator；另一类是基于机器学习的方法，包括集成学习相关算法和基于神经网络的方法，例如PeptideART、MS2PIP、MS2PBPI和pDeep等. 本文对这两大类方法进行了整理和综述，并简要指出了目前理论谱图预测方法存在的一些不足，展望了未来的发展方向.

关键词 质谱，蛋白质组学，移动质子模型，机器学习，深度学习
中图分类号 Q51，TP39 **DOI**：10.16476/j.pibb.2018.0201

基于串联质谱技术的蛋白质组学已经成为生命科学领域的重要技术. 鸟枪法蛋白质鉴定是蛋白质组学的一个关键环节，其流程如图1所示：首先实验人员将待鉴定的蛋白质样品进行酶切，产生肽段；接着对肽段进行色谱-质谱实验，生成对应的实验串联质谱图（以下简称实验谱图）；然后对实验谱图进行鉴定，得到肽段序列；最后根据肽段序列进行推断，得到蛋白质样品中存在的蛋白质序列. 在整个流程中，对实验谱图进行准确鉴定是整个过程中最重要且最困难的环节. 目前，谱图鉴定方法包括谱图库搜索、蛋白质序列库搜索和从头测序.

当目标蛋白质样品存在完整的谱图数据库时，谱图库搜索是最为理想的蛋白质鉴定方式. 此时蛋白质搜索引擎将实验谱图与谱图库中真实肽段的谱图进行比对，从而达到鉴定肽段的目的[1-2]. 虽然SRMAtlas[3]、PeptideAtlas[4]、ProteomicsDB[5-6]和HumanProteomeMap[7]等项目正在构建蛋白质谱图数据库，然而有报道称ProteomicsDB和HumanProteomeMap所构建的谱图库存在遗漏或者错误[8-9]，说明完整而又准确的谱图数据库的建立较为困难，因此目前谱图库搜索还没有成熟的解决方案.

因为谱图库的缺乏，所以蛋白质序列库搜索成为了蛋白质组学的首选方案，它也是发展得最为成熟的鉴定方法. 早期的蛋白质序列数据库搜索引擎包括Sequest[10]、Mascot[11]、X! Tandem[12]、OMMAS[13]、SCOPE[14]、ProbID[15]等；后来也相继出现了更多相关的搜索引擎，例如pFind[16]、MS-GFDB[17]、PEAKS DB[18]等.

如果某些蛋白质样品的序列库不完整，则需要利用从头测序方法直接从谱图中推断出肽段序列，常用的从头测序算法包括pNovo[19]、Peaks[20]和PepNovo[21].

无论是蛋白质序列数据库搜索还是从头测序，

* 国家重点研究发展计划（2016YFB1000605）资助项目.
** 通讯联系人. Tel：18600236326
张知非. E-mail：zhifeiz@ccmu.edu.cn
詹剑锋. E-mail：zhanjianfeng@ict.ac.cn
收稿日期：2018-07-18，接受日期：2018-12-10

Fig. 1　The workflow of protein identification in shotgun proteomics

图1　鸟枪法蛋白质鉴定流程

搜索引擎都需要根据候选的肽段序列，利用人工经验或者数学模型预测其理论谱图，然后计算理论谱图与实验谱图的相似度，用以确定此实验谱图是否来自该肽段序列. 由此可见，提高理论谱图预测的准确性是提高肽段鉴定正确性的重要部分.

在一张串联质谱图中，谱峰信号的信息有两个维度——荷质比和强度. 荷质比即离子质量与电荷的比值，代表着碎片离子质量信息；强度则代表着碎片离子的数量信息，它们与肽键键能和外部的碎裂条件等信息高度相关. 碎片离子的荷质比主要受到碎裂模式的影响. 蛋白质组学中常用的碎裂模式主要有基于碰撞碎裂的碰撞诱导裂解（collision-induced dissociation，CID）和高能碰撞裂解（higher-energy collision dissociation，HCD），它们主要断裂氨基酸之间连接的肽键，产生一系列b/y离子；另外还有电子捕获裂解（electron capture dissociation，ECD）和电子转运裂解（electron transfer dissociation，ETD），它们主要断裂氨基酸之间的N — C_{α}键，产生一系列c/z离子. 目前还出现了CID/HCD和ETD两种碰撞方式结合的新碎裂模式ETciD / EThcD（electron - transfer collision - induced/higher-energy collision dissociation）[22]，它先通过ETD，后通过CID/HCD对肽段进行碎裂. 虽然碎裂模式多样，但在已知肽段序列和碎裂模式的情况下，碎片离子的荷质比可以直接通过氨基酸残基质量计算得到，然而强度信息却暂时无法直接用数学物理公式进行计算，只能通过模型进行预测. 目前存在两种预测谱图强度的方法：基于统计物理模型的预测方法和基于机器学习模型的预测方法. 基于统计物理模型的方法主要利用移动质子模型（也被认为是一种动力学模型），已有的算法包括MassAnalyzer[23-24]和MS-Simulator[25]；基于机器学习的方法主要利用现有质谱数据及其鉴定结果，通过机器学习方法学到肽段与谱图之间的数学关系，从而预测肽段的理论谱图，已有的算法包括基于前馈神经网络的PeptideART[26-27]、基于梯度提升回归树（gradient boosting tree regression，GBRT）的MS2PBPI[28]和基于深度学习的

pDeep[29]等. 相关软件及其下载链接如表1所示.

我们对近几年发展出的理论谱图预测方法进行分类，结果如图2所示.

本文就近几年发展出的理论二级谱图预测方法进行综述：首先介绍描述肽段碎裂规律的移动质子模型；接着介绍基于移动质子模型的预测方法；然后介绍基于机器学习的预测方法；最后对现有的蛋白质组学中理论谱图预测方法进行总结并且对未来的发展进行展望.

Table 1 Information of software tools for the MS/MS spectrum prediction

表1 串联质谱图预测软件相关信息

软件名称	论文发表时间	下载地址	软件最新发布时间	获取途径	支持碎裂类型
MassAnalyzer	2004	https：//www.thermofisher.com/order/catalog/product/OPTON-30416	集成在BioPharma Finder软件中	收费	CID/HCD
MS-Simulator	2012	http：//www.bioinfo.org.cn/OpenMS-Simulator/	2015-08-15	开源	2价态的CID/HCD
PeptideART	2010	https：//sourceforge.net/projects/peptideart/	2013-04-15	开源	CID
MS2PIP	2013	https：//compomics.com/ms2pip/	2013-06-12	开源	CID
MS2PBPI	2014	https：//code.google.com/archive/p/ms2pbpi/	2014-03-13	开源	CID
pDeep	2017	http：//pfind.ict.ac.cn/download/pDeep.zip	2017-09-29	开源	HCD/ETD/EThcD

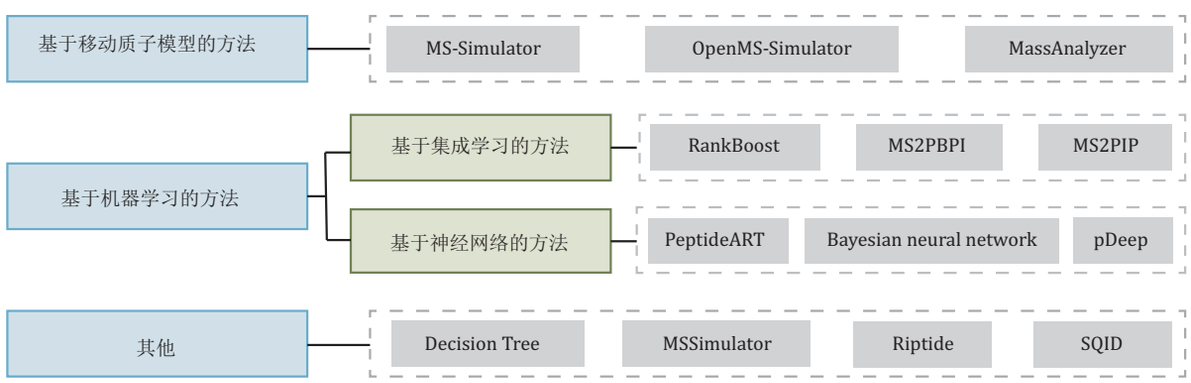

Fig. 2 The overview of spectrum prediction algorithms

图2 理论谱图预测方法分类

1 肽段碎裂的移动质子模型

在描述肽段碎裂规律的物理模型中，应用最广泛的是移动质子模型，该模型主要针对CID和HCD[30]. 此外，研究人员还提出了其他模型，比如PIC（pathways in competition）模型[31]，其核心依然是移动质子模型. 所以本文主要介绍移动质子模型.

移动质子模型认为，在带电的肽段即母离子中，如果氨基酸碱性强，则该氨基酸具备吸引质子的能力，碱性越强则代表吸引能力越强. 碱性氨基酸主要有R、H、K，其中R氨基酸碱性最强，H、K的碱性较弱. 该模型根据母离子中的碱性氨基酸数目和该母离子的质子数目（即电荷数）之差，将肽段碎裂规律分为以下两种情况：

a. 质子数目大于碱性氨基酸的数目，每个碱性氨基酸能且仅能锁住一个质子，所以还存在游离在外的质子. 游离的质子会导致其所在之处的化学键更脆弱，从而在碰撞过程中更容易断裂. 此时，肽段的断裂遵循质子导向（charge-directed）碎裂原则.

b. 质子数目小于或等于碱性氨基酸的数目，则每个质子都被一个碱性氨基酸锁定. 此时，肽段的断裂遵循质子远离（charge-remote）碎裂原则，即质子所在的地方不容易断裂，离质子远的地方更容易断裂. 本文将这种碎裂方式也称为“非寻常碎

裂”. 在锁定质子的氨基酸处，如果此位置要产生断裂，则需要足够的能量将这个位置的质子移开，由于碱性氨基酸对质子的锁定能力很强，所以在离质子越近的位置断裂就需要越高的能量，离质子越远的位置相对更容易产生断裂.

质子导向（charge-directed）和质子远离（charge-remote）是肽段结构碎裂的主要途径. 真实情况是两种碎裂途径的混合，而且会包括一些次级的途径产生碎片离子. 相比于质子导向和质子远离途径，次级途径产生的离子数量少，而且产生过程随机，所以对次级途径的研究难度非常大，目前研究人员暂时还没有对次级途径产生的离子进行精密研究.

Wysocki等[30]对非寻常碎裂进行了特别的实验验证. 他们发现，当没有游离的质子时：a. 若肽段中同时含有酸性氨基酸（如氨基酸D或E）和弱碱性氨基酸（如氨基酸H），则碱性氨基酸会诱导酸性氨基酸失去质子，导致酸性氨基酸容易发生碎裂；b. 如果弱碱性氨基酸带质子，带电的键会变得脆弱从而更容易断裂，此时非寻常碎裂发生在弱碱性氨基酸上.

虽然移动质子模型一直都没有得到数学物理学方面的严谨证明，但是它能够对部分肽段的碎裂提供较好的解释. 另外一些研究人员通过一些物理假设，将移动质子模型应用到理论谱图的预测上，取得了不错的预测结果[23-25]. 基于移动质子模型的算法包括MassAnalyzer和MS-Simulator. 值得注意的是，在表示预测谱图与实际谱图的相似度时，绝大多数算法，包括MassAnalyzer和MS-Simulator，使用的都是预测谱图与实际谱图的皮尔逊相关系数（Pearson correlation coefficient，PCC）.

2　基于移动质子模型的预测方法

2.1　MassAnalyzer

MassAnalyzer[23-24]是首个将移动质子模型用到理论谱图预测问题上的算法，它提出了一个反应动力学模型来表示肽段的碎裂过程. 假设母离子P有可能产生$F_1, F_2, \ldots, F_n$这n种碎片离子，其对应的反应速率为$k_1, k_2, \ldots, k_n$，则碎裂的动力学过程可以表示为：$[\mathrm{P}]_t = [\mathrm{P}]_0 \exp(-k_{\mathrm{total}} t)$，其中$[\mathrm{P}]_t$表示在$t$时刻还剩下的母离子P的总数，说明$[\mathrm{P}]_0 - [\mathrm{P}]_t$个母离子都被碎裂成了碎片，$k_{\mathrm{total}} = \sum_{i=1}^{n} k_i$. 则在$t$时刻，碎片离子$F_i$的丰度为$[F_i]_t = \frac{k_i}{k_{\mathrm{total}}}\left([\mathrm{P}]_0 - [\mathrm{P}]_t\right) = \frac{k_i}{k_{\mathrm{total}}}[\mathrm{P}]_0\left(1 - \exp\left(-k_{total}\Delta t\right)\right)$. 所以只要能求出反应速率$k$，就能求出碎片离子的丰度. 为了求解$k$，MassAnalyzer基于移动质子模型，提出了9个数学假设以便对谱图预测问题进行建模：

a. 存在有效温度T_{eff}可以用来表示离子的内能. 当母离子处于激发状态时，T_{eff}和母离子的电荷状态、质量和碰撞能量呈线性关系.

b. 所有肽段的碎裂过程可以分为两个阶段：（1）分子内的质子转移阶段；（2）碎裂阶段. 相对于碎裂阶段，分子内的质子转移是一个很快的进程. 所以对于质子导向（charge-directed）的情况，$k = pk'$，即肽段碎裂过程的速率常数k等于碎裂端的质子密度p乘以碎裂的速率常数k'. 对于质子远离（charge-remote）的情况，$k = (1-p)k'$.

c. 碰撞过程中的能量交换比分解过程的速率快很多. 在快速能量交换的时刻，对于碎裂阶段的速率常数k'可以通过阿仑尼乌斯方程（表示温度对离子运行速率的影响，来自热力物理学，也称Arrhenius方程）表示，即$k' = A\exp\left(-E_a/RT_{\mathrm{eff}}\right)$，其中$A$是频率因子，$E_a$是激发能量，$R$是气体常数.

d. 碎裂产生的任何离子的温度T_{eff}都随着缓冲气体的温度T_0指数下降，即$T_{\mathrm{eff}} = \left(T_{\mathrm{precursor}} - T_0\right)\exp\left(-r_c t\right) + T_0$，其中$T_{\mathrm{precursor}}$是母离子的有效温度，$t$是母离子碎裂过程的时间. 冷却速率$r_c$被假定只和离子的质量$M$相关，即$r_c = r_c^0\left(M/1000\right)^c$，其中$c$是个常数，$r_c^0$表示一个质量为1 000 u的离子的冷却速率.

e. 对于所有的碎裂路径，假设c中的频率因子A都是相同的.

f. 肽段某个位置的断裂所需要的激活能量取决于近邻氨基酸的构成，而且近邻氨基酸对该位置的影响是互相叠加的.

g. 碎裂位置的气相碱性（gas-phase basicity，GB）取决于近邻两个氨基酸的构成，而且该影响是叠加的. 在计算GB时，N端、C端和中间的氨基酸不作区分.

h. 不同侧链的气相碱性GB是固定的常数，与环境无关.

i. 对于离子的质子脱落，所需要的激发能量与该离子的表面气相碱性呈线性相关.

在满足这9个假定条件的前提下，MassAnalyzer总共考虑了11种肽段碎裂路径. 模型用了236个参数，这些参数值都由实际谱图数据统计得到. 通过比较实验谱图和预测的理论谱图，模型在训练数据集上的结果平均PCC为0.71，标准差为0.1；在测试数据集上的结果平均PCC为0.73，标准差为0.08. 最初模型仅能预测1价和2价离子，后来模型通过更新质子分布的假设和离子丰度的假设，可以预测3价及以上的离子[24].

MassAnalyzer首次将移动质子模型应用到理论谱图预测问题上，而且实验结果从一方面证明了模型的有效性，这为移动质子模型在理论谱图预测问题上的应用推动做出了很大的贡献. 该方法成为了最常用的谱图预测方法之一，已经被集成到商业软件BioPharma Finder中. 但是MassAnalyzer使用的假设条件比较多，而且仅考虑11种肽段碎裂路径，在一定程度上限制了MassAnalyzer的扩展性.

2.2 MS-Simulator

MS-Simulator[25]也是基于移动质子模型的理论谱图预测方法，相比于其他方法直接预测碎片离子的强度值，MS-Simulator仅预测某一碎裂位置的相邻碎片离子的强度比值. 当离子的强度比值预测完成后，只要设置其中一个碎片离子，比如y_1离子的强度为10 000，则其他离子的强度值可以通过比值计算得到. 相对于MassAnalyzer，MS-Simulator采用了更加简化的4个假设：

a. 碎裂产生的离子强度y_i与该离子的质子化强度P_i成正比，即$y_i=\alpha\cdot P_i$，其中α是归一化参数，P_i表示肽段中氨基酸被质子化的概率，即获得质子的概率；

b. 氨基酸质子化概率服从Boltzmann分布，即$P_i=\exp(\beta\cdot E_i)$，其中$\beta=\frac{-1}{R\cdot T_{\text{eff}}}$，$R$是亲和常数，$T_{\text{eff}}$是影响碎裂的有效温度. 在仪器相同且碎裂能量相等的情况下，T_{eff}是个常数，所以β也是常数. E_i为离子能量强度，表示氨基酸残基被质子化之后所处的能量等级；

c. 某位置的能量强度取决于周围氨基酸的能量强度对该位置的影响；

d. 离某位置距离远的氨基酸对于该位置能量强度的影响几乎为零.

综合假设a和b可得出$\log(\frac{y_i}{y_{i+1}})=\beta\cdot(E_i-E_{i+1})$，即相邻离子强度比值取决于离子之间能量强度$E_i$的差值及其常数系数$\beta$. 综合假设c和d可以建立能量强度$E_i$与相邻氨基酸的数学关系. 为了预测某一碎裂位置的相邻碎片离子的强度比值，模型只需要学习到附近氨基酸对离子的能量强度的影响. MS-Simulator通过较大规模的实验谱图来学习不同氨基酸在不同距离下的能量强度值. 实验结果表明，MS-Simulator的PCC均值可以达到0.92. 同等情况下，MassAnalyzer的PCC只有0.79.

但是，MS-Simulator模型的使用有一定的限制，例如模型只能用于CID的数据集，而且只能预测2价母离子的1价y离子的谱峰强度. 这让此模型的适用性大大减弱. OpenMS-Simulator[32]是MS-Simulator的改进模型，它在MS-Simulator中进行了少许扩展，使得模型不仅在CID数据上适用，也在HCD数据上适用.

MS-Simulator在MassAnalyzer的基础上，简化了大部分假设，而且性能相对于MassAnalyzer有了很大提升. 虽然目前只能预测2价母离子的1价y离子，相信此方法可以推广到其他离子类型上. 然而，基于移动质子模型的方法都存在一个缺陷，即方法必须建立在一些假设的基础上，限制了方法的灵活性. 所以对移动质子模型或者肽段碎裂的物理模型本身，还存在更多的探索空间.

3 基于机器学习的预测方法

由于肽段碎裂的物理过程非常复杂，至今我们都很难用数学物理学方法对这个过程进行准确建模. 下面我们将讨论一种不过分依赖物理方法的模型——机器学习模型. 蛋白质组学技术的快速发展，给我们积累了大量高质量的质谱数据，因而为机器学习的有效训练提供了可能. 基于机器学习的谱图预测模型的核心是学习一个函数$y=f(x)$，其中x为肽段，y为预测的碎片离子集合或向量. 一个碎片离子的信息包括离子的质量和强度，在已知肽段的情况下，离子的质量可以根据氨基酸残基质量计算得到，所以谱图预测的核心问题是准确预测碎片离子的强度. 实际上，前面我们描述的物理模型就是通过数学物理方法推导出函数f的形式，而机器学习方法则是利用机器学习模型结合基于大规模数据的训练，以代替数学物理方法得到有效的

预测模型. 很多研究人员尝试使用机器学习方法来解决理论谱图预测问题. 基于机器学习的谱图预测方法，目前可分为两大类，其中第一类为集成学习相关的算法，比如随机森林模型MS2PIP[33]、基于Boosting算法的MS2PBPI[28]模型和基于RankBoost的排序模型[34]等；第二类为基于神经网络的方法，包括PeptideART[26-27]、Bayesian神经网络[35]和深度学习模型pDeep[29]等. 下面我们主要介绍此两大类基于机器学习的谱图预测方法.

3.1 基于集成学习的相关模型

基于集成学习的方法，主要特性体现为需要人为给定模型的特征，之后利用大数据学习到模型特征的权重. 基于集成学习的相关算法包括基于随机森林模型的MS2PIP、基于Boosting模型的MS2PBPI和基于RankBoost的排序模型.

3.1.1 基于随机森林的MS2PIP模型

MS2PIP是基于随机森林的模型[33]. 随机森林是比较常用的集成学习算法[36]，已经在基于Kinect等设备的人体姿态识别问题中得到了广泛应用[37]. 在预测谱图中谱峰的强度值时，MS2PIP根据离子带电荷数和肽段长度将训练数据中的肽段分为42类，每类肽段都提取了一些关键特征进行表示，包括某位置是否为N端、肽段的质量数、碎片离子质量数、肽段中氨基酸的平均化学性质、碎片离子中氨基酸平均化学性质、肽段中氨基酸数目，其中氨基酸的化学性质包含4类，分别为碱性、疏水性、螺旋性和等电点.

MS2PIP针对分类后的肽段分别训练模型，最终得到42个模型的集合. MS2PIP与基于神经网络方法的PeptideART模型（PeptideART模型的详细介绍见**3.2.1**节）进行实验对比分析，在提供的5个测试数据集上，MS2PIP得到的PCC结果均好于PeptideART. 但是对于长度更长的肽段，两种方法的预测精度均会下降.

MS2PIP文章中指出其最大的贡献是根据电荷状态、肽段长度对数据进行分类，这使得模型更容易学习到一类数据中的规律，这也是MS2PIP模型能够提升精确度的关键. 但这同时也带来了一定的复杂性. 例如在对模型进行训练时，每个模型都需要大量可用的数据集，但是某些肽段却无法获得大量数据，例如一些翻译后修饰的肽段，所以这在一定程度上限制了模型的扩展性.

3.1.2 基于Boosting的模型MS2PBPI

除了随机森林，Boosting也是一种常用的集成学习算法[38]，MS2PBPI就是基于Boosting的理论谱图预测方法[28].

通过收集所有可能影响碎裂的表示特征，MS2PBPI使用33个特征表示肽段的碎片离子，包括肽段长度，肽段中N、C端的氨基酸，肽段中P、D、E氨基酸和碱性氨基酸的个数，肽段中是否有移动质子，碎片离子的荷质比，碎裂位点与N、C端的距离，碎片离子的平均气相碱性，N端残基的螺旋性、疏水性、气相碱性、质子吸引力等. 对于修饰肽段的表示，模型额外增加6个特征，包括肽段中各种修饰的数目，即氨基酸C甲烷还原烷基化的数目、氨基酸M氧化的数目、N端乙酰化的数目、N端pyro-Glu和pyro-Gln的数目、脱酰胺的数目、N端和C端的残基中修饰的总数目.

Boosting算法的目标是将多个弱预测器集成并提升成为强预测器，因此MS2PBPI首先需要构建弱预测器. 为了构建弱预测器，MS2PBPI首先将数据进行分类，其中非修饰的数据被构建为60棵分类二叉树，修饰和磷酸化的数据被构建为69棵分类二叉树. 训练数据包含66万条谱图，其中包括磷酸化的肽段数据12万张谱图. 通过使用分类二叉树，数据被划分为65 505个分类进行弱预测器的构建. 之后，细分的数据分别通过Boosting算法构建集成模型，最终MS2PBPI得到2 420个强预测器. 这些模型不仅能在保证良好的预测准确率的情况下预测非修饰肽段的谱图，还能预测修饰肽段的谱图.

MS2PBPI首次构建了磷酸化肽段的预测模型，不过由于它需要对数据进行大量划分并且需要构建成百上千个模型，这使得模型构建的复杂性大大增加.

3.1.3 基于RankBoost的排序模型

由于现阶段肽段离子碎裂的物理过程并不准确地为人所知，所以提高理论谱图预测的精确度一直是我们追求的目标. 相比于直接预测谱图中所有谱峰的强度值，在2009年Frank[34]提出仅预测所有谱峰强度的相对排序，这极大地降低了模型预测的难度. 模型使用Freund等[39]提出的排序学习模型——RankBoost进行构建. RankBoost也是以Boosting算法为核心实现的，其基本思想为：首先用训练数据训练一个弱的排序器，然后根据这个排序器的排序结果对下一次训练数据进行权重调整，其中没有被正确排序的数据会被赋予更大的权重. 重复这个步骤，一直到训练好常数T个排序器，最

后将这T个排序器进行加权合并，最终得到强的排序模型.

在数据处理时，数据根据离子带电数目、母离子质量和肽段带电荷数与碱性氨基酸数目相对差值被分为39类. 在数据特征表示时，每个肽段仅考虑4种丰度最高的离子类型，每种离子类型使用超过200个特征表示，这些特征分为3类：a. 最高强度位置特征，即最高强度离子的质量在所有离子质量中的位置信息；b. 相连氨基酸特征，即断裂位点前后3个氨基酸的信息；c. 肽段组成特征，比如与肽段的N端和C端的距离信息.

在测试结果中，RankBoost对于1价母离子预测的PCC为0.68、2价母离子为0.69、3价母离子为0.47. 因为该模型仅预测肽段谱峰的相对排序，与之前描述的预测模型结果表示方法不同，所以导致其PCC计算结果比较低.

Frank将预测理论谱图的谱峰强度问题转化成一个排序问题，思路比较新颖，但他对数据分类太细使模型的适用性不强，需要训练多个模型才能覆盖所需要预测的肽段种类，这也是上述所有集成学习相关算法弊端之一.

3.2 基于神经网络的相关模型

基于集成学习的传统方法取得了一定的成效，但随着机器学习的深度发展，研究人员也将神经网络方法应用到理论谱图预测问题上. 相比于传统方法，神经网络近年来被证明可以不用人为设计特征，只需要大量数据，就可以直接让模型在数据中学习到所需要的参数和特征[40].

3.2.1 基于神经网络的PeptideART模型

早在2006年，Arnold等[27]就将神经网络用于理论谱图预测问题上，但是由于当时软硬件水平的限制，他们只能使用浅层的神经网络，该模型称为PeptideART.

模型使用的特征包括氨基酸组成、碎裂位点、碎裂位点前后两个氨基酸、N端C端氨基酸、母离子质量、子离子质量、气象碱度、螺旋性、疏水性、等电点等. 选取特征之后，Arnold等使用两层前馈神经网络构建PeptideART，目标是预测2价和3价离子. 模型的数据根据离子类型被分类，之后被分开训练，最终实验得到40个模型.

经过几年发展，PeptideART模型在2010年被Li等[26]重新评估. 他们发现随着实验使用的数据集不同，PeptideART的实验结果准确性变化很大. 对于低精度CID数据集，使用相同物种的相同肽段产生的数据进行实验，PeptideART实验结果的相似度PCC为0.85，但是使用不同物种的相同肽段进行实验，实验结果PCC仅为0.70. 原因可能是低精度数据受到肽段的其他信号和噪音信号干扰较大，所以较难预测.

即使不同数据集导致预测结果准确性下降，PeptideART的准确性PCC结果依然在0.7以上. 在与MassAnalyzer的比较实验中，PeptideART预测理论谱图的准确性依然有一定程度的提升.

PeptideART探索了神经网络在预测理论谱图问题上的应用，但是使用不同物种的相同肽段进行实验时，模型的精度波动较大，说明早期的低精度数据并不适合进行谱图预测. 随着仪器的发展，现阶段涌现了很多高精度数据，在这些数据和新仪器上，谱图预测问题被验证是可重复的.

3.2.2 基于贝叶斯神经网络的模型

与PeptideART相似，Zhou等[35]使用了贝叶斯神经网络来构建模型. 在特征选取方面，Zhou等选择了在已有研究中35个已经确定对肽段碎裂有影响的特征[41]，包括碎裂位点距离N、C端和肽段中心点的距离，碎裂位点是否位于边缘，N、C端碱性、螺旋性和疏水性，整个肽段的碱性、螺旋性、疏水性，N端和C端的等电点值，肽段长度，肽段中碱性氨基酸数目，肽段质量，碎裂y离子质量等. 该文的目标在于发现CID数据集中对肽段碎裂影响最大的特征. 通过使用Kapp等[42]提出的RMP（relative mobile proton）假说，文章将肽段分为3类，即质子可移动肽段、质子部分移动肽段和质子不可移动肽段. 其中质子可移动肽段指肽段中碱性氨基酸的数目，即R、H、K氨基酸总数少于肽段的质子数；质子不可移动肽段指肽段中R氨基酸数目大于或者等于质子数；其他则属于质子部分移动肽段，即R氨基酸数目少于质子数但是碱性氨基酸的数目大于质子数. 该文认为这三类肽段的碎裂规律并不相同，所以应该进行分别研究.

在进一步探索上述特征在表示肽段碎裂的重要性时，该方法使用3层前馈神经网络和35个特征来构建贝叶斯神经模型，数据方面该方法只选择2价离子进行实验. 通过使用自动关联决定（automatic relevance determination，ARD）技术[43-44]定义非关联参数，贝叶斯神经网络模型可以输出每个特征的非关联参数，即表征该特征的非相关性，评分越高即表示该特征对于肽段碎裂的表示越不重要. 研究结果表明质子可移动肽段和质子部分移动肽段的

非关联参数的排列类似，但是与质子不可移动肽段的非关联参数排列很不相同. 例如N、C端的平均疏水性和N、C端的疏水性差异对于非移动肽段没有重要的影响，但是对于移动肽段和部分移动肽段有重要的影响. 而有一些特征对于所有肽段表示都有重要的影响，例如N、C端的碱性，碎裂位点距离N端、C端、肽段中心的距离，碎裂位点是否位于两端，碎裂位置左右两端的氨基酸等. 这也证实了质子移动模型当中以R氨基酸数目和质子数目差来区分肽段的有效性.

对于质子不可移动肽段，实验中通过对参数进行重要性排序，最不重要的特征被逐个删除，每个特征被删除后，模型进行重新构建，用以验证模型结果的精确性并没有显著下降. 当第23个特征被删除时，模型的错误率会显著提高，所以最终22个不重要的特征被删除，13个特征被保留，这使得用于表征肽段的特征数尽可能少，同时对于实验结果准确度没有较大的影响. 对于质子可移动肽段和质子部分移动肽段，作者没有进行这部分的实验探究.

对于质子不可移动肽段数据，13个特征被用于构建3层前馈神经网络模型，即得到贝叶斯神经网络模型. 模型与MassAnalyzer进行实验结果的对比，预测结果精确度相差无几.

贝叶斯网络在保证谱图预测精度的前提下，尽量减少所用特征，以发现对肽段碎裂影响最大的几个特征，对于探索肽段碎裂特性有深远的影响.

3.2.3　基于深度学习的pDeep模型

浅层神经网络在理论谱图预测问题上有很好的应用，但是准确率还不够高. pDeep首次将深度神经网络应用到理论谱图预测问题上[29]. 由于肽段中的化学键的碎裂是和肽段的整个氨基酸结构相关的. 传统的浅层神经网络认为样本独立同分布，所以将氨基酸的各个碎裂位置进行独立处理，然而此做法导致模型学习不到氨基酸之间的相互影响. 为了建模碎裂过程中氨基酸的相互影响，pDeep采用了深度学习中的双向长短时记忆网络（bidirectional long short-term memory，BiLSTM）. 因为长短时记忆网络（long short-term memory，LSTM）具有长短时记忆功能，它可以学习到N端或者C端氨基酸对于当前化学键碎裂的影响. BiLSTM是双向的LSTM网络，所以它可以同时学习到N端和C端的氨基酸对某化学键断裂的综合影响. 关于LSTM网络的思想和算法可以参考文献［45］.

pDeep使用BiLSTM网络进行模型构建，方法如图3所示. 在考虑模型输入特征时，pDeep考虑了碎裂位置前后氨基酸信息、N端和C端的氨基酸信息和电荷数等信息. 与其他谱图预测模型不同的是，pDeep并没有对数据集进行任何分类，所有数据都用统一的模型进行构建，并且利用深度学习的建模能力完成模型的学习.

在探究深度学习在理论谱图预测方面的应用上时，我们首先通过跨实验室和跨物种的实验证明，理论谱图预测问题是一个可重复性的问题，这是使用机器学习方法的基础. 在进行同类软件对比时，我们不仅将模型与基于移动质子模型的MassAnalyzer、MS-Simulator进行对比，而且与基于浅层神经网络的PeptideART进行了对比，pDeep都实现了更高的预测准确度.

由于pDeep不依赖任何前提假设，所以模型在理论谱图预测问题上具有很好的可扩展性. pDeep不仅在HCD数据集上可以进行良好的预测，在ETD和EThcD数据集上的预测结果也可以得到超过0.9的PCC.

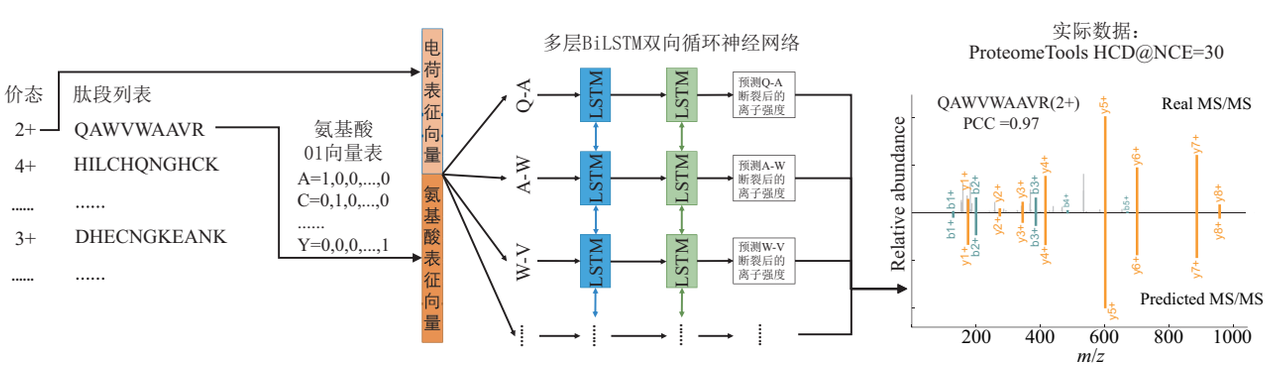

Fig. 3　The BiLSTM-based pDeep model

图3　基于BiLSTM双向循环神经网络的pDeep模型

pDeep还可以用来区别极度相似的氨基酸，比如“GG”和“N”，它们的氨基酸残基质量完全相同，所以使用荷质比信息很难将两者区分开. 通过使用pDeep模型，我们发现氨基酸“GG”和“N”的区分率可以达到90%以上. 对于“AG”和“Q”，pDeep也能达到90%以上的区分率.

在模型训练的过程中，深度学习可以自主学习到肽段中氨基酸组成对碎裂的影响. 通过比较pDeep模型的中间层信息，我们可以得到氨基酸N端和C端肽键碎裂性质的相似性关系，其热力图如图4所示. 从图4可以看出，pDeep学习到的相似性关系与氨基酸的化学结构有一定的关联，比如氨基酸F和Y都含有苯环，所以二者的碎裂性质非常接近. 从图4中可以看出深度学习可以自动从数据中学习到氨基酸内部的理化性质，这也从侧面验证了深度学习的有效性.

pDeep首次将深度学习应用到理论谱图预测问题上并得到了很好的效果，展现了深度学习在理论谱图预测问题上应用的潜力. 但是基于深度学习的模型是一个黑盒子，并不能让人全盘了解深度学习所学习到的信息，所以可解释性较弱.

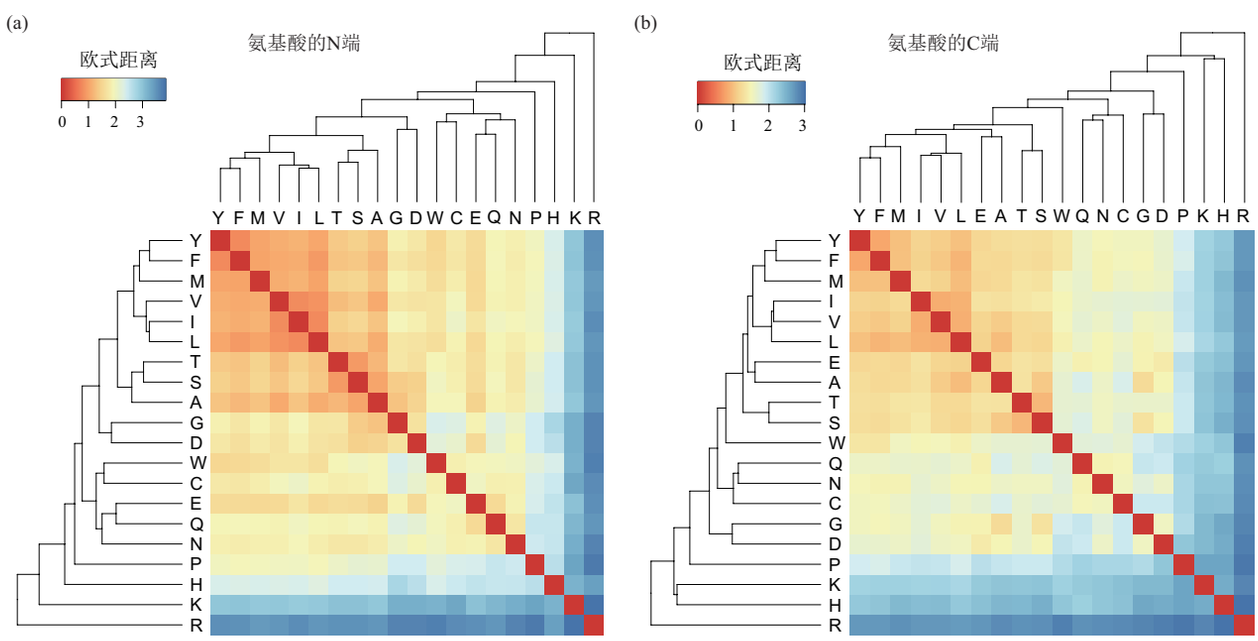

Fig. 4　Heat map of fragmentation similarities of amino acids learned by pDeep

图4　pDeep学习到的氨基酸碎裂性质的相似性热力图

（a）肽键N端的氨基酸.（b）肽键C端的氨基酸.

4　其他谱图预测相关方法

除了上述几种模型，还有一些理论谱图预测相关的工作. 这些工作并不主要用于理论谱图预测，而是用于谱图匹配打分等其他应用，但是由于涉及了部分谱图预测方面的工作，所以该文对这些方法进行简要介绍.

Elias等[41]将决策树模型应用在谱图预测问题上，该模型主要用于提升肽段鉴定过程中谱图匹配打分策略的准确性. 理论谱图预测问题在这里被视为一个分类问题，所以模型采用决策树的方法进行构建. 在使用高精度数据集训练了一个谱图预测模型后，Elias等使用相同数据中谱图鉴定结果排序的第二位谱图，即和正确谱图极相似的错误谱图，构建一个错误的谱图预测模型. 在谱图预测过程中结合这两个模型对实验结果进行谱图匹配打分. 实验证明，这两个模型对于区分实验中的假阳性结果有一定的作用.

SQID通过统计两个氨基酸之间出现离子的概率来对理论谱图进行预测，预测结果能够有效提高

肽谱匹配打分的性能[46]. Riptide也利用谱图预测来提高肽谱匹配打分的性能，不过它使用的模型是动态贝叶斯网络[47].

MSSimulator也是谱图预测相关软件，不过它的主要目标是生成基准测试谱图集[48]. 它提供了3个模型以生成模拟谱图：第一个模型由用户自己给出各个离子的强度值；第二个模型使用支持向量机（support vector machine，SVM）算法给出各个离子出现的概率；第三个模型使用支持向量回归（support vector regression，SVR）算法给出离子的强度. 由于MSSimulator本身的定位是基准测试，因此谱图预测并不是它最重要的工作.

5　总结与未来展望

肽段的理论二级谱图预测方法一直在发展，还有很多可以提升的空间. 究其原因，是由于肽段的碎裂规律太过复杂，我们暂时无法完全了解并建模其中的规律. 基于物理模型的方法和基于机器学习模型的方法对于谱图预测各有长短：物理模型注重于肽段内在碎裂规律的探索与研究，但由于无法精密地观察肽段碎裂的过程，所以模型需要一些前提假设的辅助. 其中MassAnalyzer进行了9项假设，而MS-Simulator进行了4项假设；机器学习模型更加注重预测结果的准确性，它的主要目标是从已有的数据中学习如何更准确地预测未知的数据. 计算技术的发展，尤其是深度学习技术的发展，以及大量高质量质谱数据的积累，比如人类蛋白质组的肽段以及修饰肽段的合成项目——ProteomeTools[49-50]，使机器学习模型预测的准确性有较大提高. 遗憾的是机器学习方法缺乏可解释性，所以我们无法直接将机器学习到的规律转化成肽段碎裂的物理规律. 所以，有效结合基于物理模型的方法与基于机器学习模型的方法，将是未来的一个重要的研究方向.

参　考　文　献

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

[2] Lam H, Aebersold R. Building and searching tandem mass (MS/MS) spectral libraries for peptide identification in proteomics. Nat Methods (San Diego, Calif.), 2011, **54**(4): 424-431

[3] Kusebauch U, Campbell D S, Deutsch E W, *et al*. Human SRMAtlas: a resource of targeted assays to quantify the complete human proteome. Cell, 2016, **166**(3): 766-778

[4] Farrah T, Deutsch E W, Hoopmann M R, *et al*. The state of the human proteome in 2012 as viewed through peptideAtlas. Journal of Proteome Research, 2013, **12**(1): 162-171

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

[6] Schmidt T, Samaras P, Frejno M, *et al*. ProteomicsDB. Nucleic Acids Research, 2018, **46**(D1): D1271-D1281

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

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

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

[10] Eng J K, Mccormack A L, Yates J R. An approach to correlate tandem mass-spectral data of peptides with amino-acid-sequences in a protein database. J Am Soc Mass Spectrom, 1994, **5**(11): 976-989

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

[12] Craig R, Beavis R C. TANDEM: matching proteins with tandem mass spectra. Bioinformatics, 2004, **20**(9): 1466-1467

[13] Geer L Y, Markey S P, Kowalak J A, *et al*. Open mass spectrometry search algorithm. J Proteome Res, 2004, **3**(5): 958-964

[14] Bafna V, Edwards N. SCOPE: a probabilistic model for scoring tandem mass spectra against a peptide database. Bioinformatics, 2001, **17**(Suppl 1): S13-S21

[15] Zhang N, Aebersold R, Schwikowski B. ProbID: a probabilistic algorithm to identify peptides through sequence database searching using tandem mass spectral data. Proteomics, 2002, **2**(10): 1406-1412

[16] Sun R X, Dong M Q, Song C Q, *et al*. Improved peptide identification for proteomic analysis based on comprehensive characterization of electron transfer dissociation spectra. J Proteome Res, 2010, **9**(12): 6354-6367

[17] 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. Molecular & Cellular Proteomics, 2010, **9**(12): 2840-2852

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

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

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

[21] Frank A, Pevzner P. PepNovo: *de novo* peptide sequencing *via*

probabilistic network modeling. Analytical Chemistry, 2005, **77**(4): 964-973

[22] Frese C K, Altelaar A F, van den Toorn H, *et al*. Toward full peptide sequence coverage by dual fragmentation combining electron-transfer and higher-energy collision dissociation tandem mass spectrometry. Anal Chem, 2012, **84**(22): 9668-9673

[23] Zhang Z. Prediction of low-energy collision-induced dissociation spectra of peptides. Analytical Chemistry, 2004, **76**(14): 3908-3922

[24] Zhang Z. Prediction of low-energy collision-induced dissociation spectra of peptides with three or more charges. Analytical Chemistry, 2005, **77**(19): 6364-6373

[25] Sun S, Yang F, Yang Q, *et al*. MS-simulator: predicting Y-ion intensities for peptides with two charges based on the intensity ratio of neighboring ions. Journal of Proteome Research, 2012, **11**(9): 4509-4516

[26] Li S, Arnold R J, Tang H, *et al*. On the accuracy and limits of peptide fragmentation spectrum prediction. Analytical Chemistry, 2011, **83**(3): 790-796

[27] Arnold R J, Jayasankar N, Aggarwal D, *et al*. A machine learning approach to predicting peptide fragmentation spectra. Pacific Symposium on Biocomputing, 2006: 219-230

[28] Dong N P, Liang Y Z, Xu Q S, *et al*. Prediction of peptide fragment ion mass spectra by data mining techniques. Analytical Chemistry, 2014, **86**(15): 7446-7454

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

[30] Wysocki V H, Tsaprailis G, Smith L L, *et al*. Mobile and localized protons: a framework for understanding peptide dissociation. Journal of Mass Spectrometry, 2001, **35**(12): 1399-1406

[31] Paizs B, Suhai S. Fragmentation pathways of protonated peptides. Mass Spectrom Review, 2004, **24**(4): 508-548

[32] Wang Y, Yang F, Wu P, *et al*. OpenMS-simulator - an open-source software for theoretical tandem mass spectrum prediction. BMC Bioinformatics, 2015, **16**(1): 110

[33] Degroeve S, Martens L. MS2PIP - a tool for MS/MS peak intensity prediction. Bioinformatics, 2013, **29**(24): 3199-3203

[34] Frank A M. Predicting intensity ranks of peptide fragment ions. Journal of Proteome Research, 2009, **8**(5): 2226-2240

[35] Zhou C, Bowler L D, Feng J. A machine learning approach to explore the spectra intensity pattern of peptides using tandem mass spectrometry data. BMC Bioinformatics, 2008, **9**(1): 325

[36] Breiman L. Random forests. Machine Learning, 2001, **45**(1): 5-32

[37] Shotton J, Fitzgibbon A W, Cook M, *et al*. Real-time human pose recognition in parts from single depth images. CVPR, 2011: 1297-1304

[38] Friedman J H. Greedy function approximation: a gradient boosting machine. Annals of Statistics, 2001, **29**(5): 1189-1232

[39] Freund Y, Iyer R D, Schapire R E, *et al*. An efficient boosting algorithm for combining preferences. Journal of Machine Learning Research, 2003, **4**(6): 933-969

[40] Hinton G E. Reducing the dimensionality of data with neural networks. Science, 2006, **313**(5786): 504-507

[41] Elias J E, Gibbons F D, King O D, *et al*. Intensity-based protein identification by machine learning from a library of tandem mass spectra. Nature Biotechnology, 2004, **22**(2): 214-219

[42] Kapp E A, Schütz F, Reid G E, *et al*. Mining a tandem mass spectrometry database to determine the trends and global factors influencing peptide fragmentation. Analytical Chemistry, 2003, **75**(22): 6251-6264

[43] Marshall J A. Neural networks for pattern recognition. Neural Networks, 1995, **8**(3): 493-494

[44] MacKay D J. Bayesian methods for neural networks: theory and applications. 1995,

[45] LeCun Y, Bengio Y, Hinton G E. Deep learning. Nature, 2015, **521**(7553): 436-444

[46] Li W, Ji L, Goya J, *et al*. SQID: an intensity-incorporated protein identification algorithm for tandem mass spectrometry. J Proteome Res, 2011, **10**(4): 1593-1602

[47] Klammer A A, Reynolds S M, Bilmes J A, *et al*. Modeling peptide fragmentation with dynamic Bayesian networks for peptide identification. ISMB, 2008, **24**(13): i348-i356

[48] Bielow C, Aiche S, Andreotti S, *et al*. MSSimulator: simulation of mass spectrometry data. Journal of Proteome Research, 2011, **10**(7): 2922-2929

[49] Zolg D P, Wilhelm M, Schmidt T, *et al*. Proteome Tools: systematic characterization of 21 post-translational protein modifications by LC-MS / MS using synthetic peptides. Mol Cell Proteomics, 2018, **17**(9):1850-1863

[50] Zolg D P, Wilhelm M, Schnatbaum K, *et al*. Building proteometools based on a complete synthetic human proteome. Nat Methods, 2017, **14**(3): 259-262

Trends on Methods for Prediction of Tandem Mass Spectra of Peptides*

ZHOU Xie-Xuan[1,3], REN Rui[1,3], GAO Wan-Ling[1,3], HUANG Yun-You[1,3], ZENG Wen-Feng[2,3], KONG De-Fei[1,3], HAO Tian-Shu[1,3], ZHANG Zhi-Fei[4)**], ZHAN Jian-Feng[1,3)**]

([1)] *State Key Laboratory of Computer Architecture, Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS), Beijing* 100190, China;

[2)] *Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences, Institute of Computing Techology (ICT), Chinese Academy of Sciences (CAS), Beijing* 100190, China;

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

[4)] *Capital Medical University, Beijing* 100069, China)

Abstract　Tandem mass spectrometry (MS/MS)-based proteomics has become one of the most important tools in bioscience, and researchers now pay much attention to the prediction of MS/MS spectra for protein identification and quantification. With the accumulation of massive high-quality spectrum data and the development of computing technology, quite a few new methods were emerged to solve this problem. These methods can be divided into two catagories: mobile proton model-based methods, such as MassAnalyzer and MS-Simulator; and machine learning-based methods, including traditional machine learning and deep learning, such as PeptideART, MS2PIP, MS2PBPI and pDeep. In this paper, we investigated a wide variety of corresponding methods, and briefly pointed out the deficiencies of existing software tools, and suggested the future work.

Key words　mass spectrometry, proteomics, mobile proton model, machine learning, deep learning

DOI：10.16476/j.pibb.2018.0201

---

* This work was supported by a grant from The National Key Research Program of China (2016YFB1000605).

** Corresponding author. Tel: 18600236326

ZHANG Zhi-Fei. E-mail: zhifeiz@ccmu.edu.cn

ZHAN Jian-Feng. E-mail: zhanjianfeng@ict.ac.cn

Received: July 18, 2018　Accepted: December 10, 2018