多肽色谱保留预测及其在蛋白质组学中的应用

陈 可[1,2],李翠翠[1,2],李 博[1,2,3*]

([1]中国药科大学药物分析系,南京 211198;[2]中国药科大学蛋白质化学与结构生物学重点实验室,南京 211198;
[3]中国药科大学(杭州)创新药物研究院,杭州 310018)

摘 要 基于串联质谱的蛋白质组学分析方法往往依赖于实际谱图和理论谱图的匹配打分,而大量共洗脱肽的干扰会降低多肽和蛋白的鉴定及定量的准确性。多肽保留时间预测可将多肽色谱保留行为转变为稳定独立的特征时间属性,作为多肽鉴定的辅助和验证指标,改善多肽鉴定的准确性。复杂体系中多肽色谱保留预测也对优化蛋白质组学测定条件、提高数据非依赖采集中质谱数据的检出率和重复性具有重要意义。本文针对未修饰多肽及修饰多肽常用的色谱保留预测方法(包括基于标准化索引、多肽分子模型、氨基酸残基参数和机器学习等)进行了综述,总结各种方法的原理及其特点,并对其在蛋白质组学中的应用及发展方向进行了展望。

关键词 多肽保留时间预测;多肽保留时间;数据非依赖采集;蛋白质组学

中图分类号 R917 **文献标志码** A **文章编号** 1000-5048(2021)04-0422-09

doi:10.11665/j.issn.1000-5048.20210404

引用本文 陈可,李翠翠,李博.多肽色谱保留预测及其在蛋白质组学中的应用[J].中国药科大学学报,2021,52(4):422-430.

Cite this article as: CHEN Ke, LI Cuicui, LI Bo. Peptide retention prediction algorithm and its application in proteomics[J]. *J China Pharm Univ*, 2021, 52(4): 422-430.

Peptide retention prediction algorithm and its application in proteomics

CHEN Ke[1,2], LI Cuicui[1,2], LI Bo[1,2,3*]

[1]Department of Pharmaceutical Analysis, China Pharmaceutical University, Nanjing 211198; [2]Key Laboratory of Protein Chemistry and Structural Biology, China Pharmaceutical University, Nanjing 211198; [3]Innovative Drug Research Institute, China Pharmaceutical University, Hangzhou 310018, China

Abstract Most of the proteomics analysis methods based on tandem mass spectrometry rely on the matching scoring of actual spectrum and theoretical spectrum, the interference of a large number of co-eluting peptides could cause error in the identification and quantification of peptides and proteins. Peptide retention time prediction, as a auxiliary and verification index of the peptide, can transition the chromatographic behavior into stable independent time attributes, and improve the accuracy of the peptide identification. Prediction of peptide chromatographic retention in complex systems is also of great significance for optimizing proteomics determination conditions and improving the detection rate and repeatability of mass spectrometry data in data-independent acquisition. This review focused on the chromatographic retention prediction method of unmodified peptides and modified peptides, summarizes the content, characteristics and limitations of four types of peptide retention time prediction methods based on standardized indexes, peptide molecular models, amino acid residue parameters, and machine learning, as well as their applications in proteomics, with a prospect of their future.

Key words peptide retention time prediction algorithm; peptide retention time; data independent acquisition; proteomics

目前,绝大多数蛋白质组学的分析都是采用基于串联质谱的自下而上(bottom-up)的方法,对酶解的肽段进行LC-MS分析,通过肽段的串联质谱数据鉴定蛋白质[1]。应用中色谱共洗脱是多肽

收稿日期 2021-02-09 * **通信作者** Tel:025-83271350 E-mail:libo@cpu.edu.cn

串联质谱鉴定中的常见问题，多达 50% 的肽段串联质谱（MS/MS）谱图中包含一个以上的肽[2]，所产生的丰富质谱数据使得合理的解析变得困难：一方面碎片离子会受到母离子和共洗脱肽段碎片离子的干扰，增加了数据解析的难度；另一方面很多共洗脱多肽无法被鉴定。此外，蛋白质的翻译后修饰（post-translational modifications，PTMs）增加了蛋白质及多肽的多样性，对数据分析工作带来了进一步的挑战。

多肽的色谱保留取决于色谱方法和多肽本身的性质，而多肽的性质在很大程度上是由它们的氨基酸序列决定的。因此在给定的色谱条件下，保留时间（retention time，RT）包含了多肽序列的信息[3-5]。多肽保留时间预测是将多肽色谱保留行为转变为稳定独立的特征时间属性，作为蛋白质组学中辅助和验证指标，增加靶向蛋白质组学的覆盖率[6]，或为数据非依赖采集（data independent acquisition，DIA）样品提供辅助信息，提高谱图匹配的准确性[12, 34]。

本文对未修饰多肽和修饰多肽保留时间预测的各类方法进行了综述，对各方法原理、模型、特点及其在蛋白质定性及定量中的应用进行总结，讨论了这些方法在蛋白质组学中预测完整蛋白质的可行性和准确性，并对多肽保留时间预测方法的发展方向及其应用前景进行了展望。

1　未修饰肽段的保留时间预测方法

为了充分利用色谱保留数据，已有众多多肽保留时间预测方法，见图 1。这些方法大致可以分为 4 类：基于多肽物理/化学信息的多肽分子模型法；基于标准肽数据的标准化索引法；基于每个氨基酸残基贡献的氨基酸残基参数法；基于大数据分析的机器学习法等。

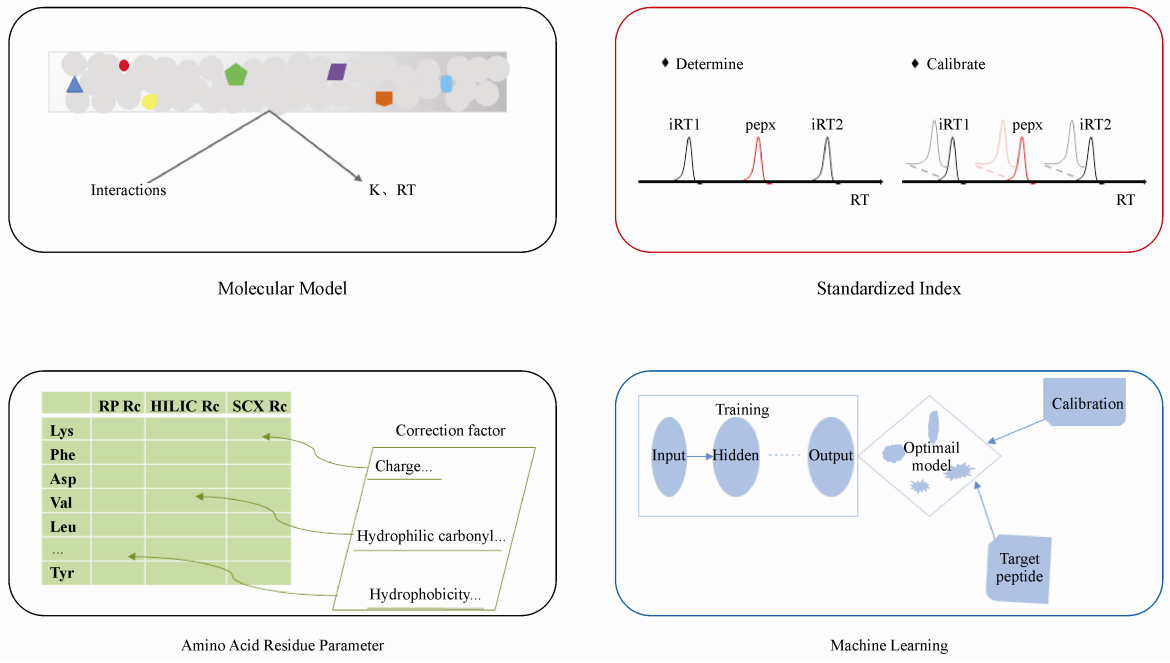

Figure 1　Four methods of peptide retention prediction: each figure illustrates the principles and characteristics of this four different methods

1.1　多肽分子模型法

在给定的色谱条件下，特定多肽的 RT 应该是恒定的，因此 RT 是化学结构依赖性参数。多肽分子模型法是通过多肽的物理化学性质即肽的结构信息或它们在分离期间的化学相互作用的信息实现多肽保留时间预测。分子模型方法偏向于对大分子进行物理建模，辅之以氨基酸残基的贡献总和进行预测，方法简便，但缺失了一些影响色谱保留的因素。

1.1.1　定量结构-保留关系（quantitative structure retention relationship，QSRR）　Kaliszan 等[7]开发的基于 QSRR 的方法使用软件计算肽序列的一系列化学特征：氨基酸残基保留时间总和的对数 $\lg Sum_{AA}$，多肽的范德瓦尔（Van der Waals，VDW）体

积的对数lg*VDWVol*,多肽的计算正辛醇-水分配系数的对数clg*P*。通过多元回归分析将其组合成预测函数,用于多肽的保留时间预测。

Le Maux等[8]则以表观亲水性、氨基酸在序列中位置、肽序列长度三者之间的函数关系,建立RT预测模型。该方法可以较为准确地预测未知短肽的氨基酸序列以及区分同源肽的保留时间。

1.1.2 临界条件生物大分子液相色谱法(liquid chromatography of biological macromolecules under critical conditions, BioLCCC) BioLCCC基于高分子统计物理学方法,利用肽链的随机游动模型及多肽分子在吸附剂孔内的空间构象对色谱分离过程进行建模,同时考虑吸附剂孔内的肽的熵和能量补偿以,及多肽和固定相之间的有效相互作用能等因素[9]。BioLCCC模型的优势在于可模拟等度或梯度条件下多肽在色谱柱上的吸附分配过程,并能直接计算出给定溶剂组成条件下多肽的保留因子[10]。

1.2 标准化索引法

标准化索引法是利用一组标准肽的保留时间建立数据库,把这些数值作为其他待测肽的RT预测的基础和标准。这样的标准肽覆盖不同的疏水性并且易于用MS检测。只需要进行一组标准肽的校正实验,就可以在后续所有不同条件的实验分析中使用其RT信息,进而改善了由于色谱系统差异导致RT数值差异很大的问题。

1.2.1 索引保留时间(indexed retention time, iRT) iRT首先由Escher等[3]提出,iRT量表的标准肽由11种不同于任何一个已知天然序列的肽构成。这是一个开放,便携和标准化的保留时间量表,它的采集窗口小,量化精度高,从而增加了LC-MS的通量和质量。目标多肽的RT是相对于标准iRT-肽的固定数值,可以跨实验室和色谱系统转移[11]。iRT精度与识别数量之间存在显著的相关性。

与多肽分子模型法相比,iRT的一系列方法应用更广泛,大大提高了蛋白质组学数据分析的检出率和准确性。但由于iRT肽数量非常有限,主要用于线性梯度条件,其精度有限。

1.2.2 高精度iRT (high-precision iRT) 为了使iRT具有更高的精度,Bruderer等[12]将iRT肽扩展到数千个,利用稳健的分段回归实现iRT和RT间的相互转换。这种高精度iRT算法能增加靶向蛋白质组学中15%的定量信息。高精度iRT的预测结果虽然能一定程度上不为色谱条件所转移,但仍需要避免操作中流动相中酸的种类及浓度变化带来的影响。

1.3 氨基酸残基参数法:从加性到序列特异性

基于残基参数的方法最初旨在预测肽段序列中每个氨基酸残基对整条肽的RT的影响。氨基酸残基的个体贡献通常被称为保留系数(retention coefficients, RC),那么整个肽的保留就是各个贡献的总和(一组RC)。在给定的色谱条件下,可以通过简单地总结(累加)组成肽的氨基酸残基的RC来估计肽的RT,这便是加性模型(additive model)。

1.3.1 加性模型 该方法最早的实例是使用一组25个短肽(胰高血糖素、生长抑素等)以及它们观察到的RT来得到序列中存在的每个氨基酸残基的保留系数[13]。使用HP 9815A计算器计算RC,并仅使用肽的氨基酸组成进行预测,而未涉及到序列中每个氨基酸的位置、空间或构象的任何信息。

随后的研究表明[14],早期的加性模型有很大的局限性,在新的色谱条件下RC需要进行重新校准;对含有50个残基的多肽需要引入肽链长度校正参数。因为即使是对于短肽,当相邻氨基酸残基不同或末端基团理化性质不同时,也可获得不同的RC[15-16]。但在这样的情况下,加性模型仍无法准确阐明吸附色谱法中肽保留的所有特征。只有非常小的肽(2~4个氨基酸残基)和没有任何二级结构才有助于实现加性模型的高预测准确性。

1.3.2 序列特异性模型 在加性模型的基础研究上,Krokhin等[4]开发了序列特异性保留计算器(sequence-specific retention calculator, SSRCalc),该算法的第1个版本使用离线HPLC-MALDI MS收集了346个胰蛋白酶肽的数据集,在加性模型的基础上进行校正,产生了两组氨基酸残基RC(一组对应于N-末端和一组对应于所有其他位置)和两组校正因子(肽长度和总疏水性)。

该算法的第2个版本便将数据集扩大至2 000,除了引入短肽的氨基酸残基的单独RC,还校正了等电点、带电肽的最近邻效应和形成螺旋结构的倾向(脯氨酸重复)。在此基础上,Elutator[2]不仅限于最近邻,进一步考虑了氨基残基的邻近效应。因为即使对于肽链中多个位置分隔开的

残基，其相互作用也具有统计学意义。

基于参数的方法的局限性就在于它们通常被优化用于预测特定色谱系统的保留时间。Dwivedi等[17]开发了二维LC系统的多肽保留预测算法。其使用了广泛的离子对和pH条件，RP（pH 10~pH 2）2D HPLC-ESI/MS系统提供了更高的一维分离效率，并增加了识别多肽的数量（约10 000个胰蛋白酶肽）。在约280 000个胰蛋白酶肽的数据集分析中，发现侧链具有N帽诱导的两亲性螺旋肽与C_{18}吸附剂的疏水作用占主导地位，其保留比预期更强[18]。于是便将描述肽的两亲性螺旋性特征（富含丙氨酸）和N帽稳定性基序（N-帽附近的N1和N2位有疏水残基天冬氨酸等）结合到SSRCalc中[19]。

在亲水相互作用液相色谱（hydrophilic interaction liquid chromatography，HILIC）系统中，携带N帽螺旋稳定基序和两亲性高螺旋的肽保留比预测值偏低，这是因为肽骨架上的亲水性羰基和酰胺基团与螺旋结构间发生氢键稳定，它决定了HILIC中独特的肽的序列依赖性行为[20]。

另一种基于SSRCalc的肽保留预测模型阳离子交换（strong cation exchange，SCX）系统的肽段分离和预测机制则是基于库仑定律驱动的肽在离子交换色谱中的静电相互作用[21]。肽的电荷越大，库仑相互作用越强，保留也就越强，碱性残基会增加肽的N末端附近的保留，酸性氨基酸则相反，疏水性氨基酸也表现出较低的保留系数。这决定了SCX中独特的肽的序列依赖性行为。

由此也能看出，对于不同的实验条件，它们的预测结果力就会发生偏差，需要引入特定的参数进行校正才能获得良好的相关性。SSRCalc是目前使用最广泛的基于参数的保留时间预测器，可以说是该领域的基准工具，也是最准确的保留时间预测模型之一。在肽的电荷、长度、疏水性、二级结构、螺旋结构，氨基酸的个体保留和相对于肽末端的位置乃至不同色谱系统等方面的优化，SSRCalc已经取得了较大进展。

1.4　机器学习法

利用人工智能的机器学习法也被用于多肽保留时间预测。方法利用计算机算法从已知的输入数据中获得信息，输出数值，进行训练。根据训练中获得的输入输出数据建立已知参数模型，对目标肽段的RT进行预测。基于机器学习的RT预测方法可以分为两大类：传统的机器学习方法和深度学习方法。机器学习方法又分为两个子类：一类为人工神经网络（artificial neural networks，ANN）[22-23]，另一类是支持向量回归（support vector regression，SVR）算法[5, 24]。

1.4.1　人工神经网络 (ANN)

最初ANN以20个氨基酸残基的组成为基础，由20个输入节点、2个隐含节点和1个输出节点组成[22]。使用约7 000个已知RT的训练肽进行网络训练，并利用来自于另一微生物种的约5 200个肽（多达54个氨基酸残基）进行鉴定评估，结合遗传算法优化线性方程系数以进行时间和梯度斜率校正，将肽保留数据归一化到一定个范围（0~1），从而将肽RT的重现性误差缩小至1%。在后续对该方法的改进中采用由1 052个输入节点、24个隐含节点和1个输出节点组成的ANN结构，同时编码了氨基酸位置，肽长度和疏水性，最近邻氨基酸以及肽的二级结构（螺旋、片状、卷曲）等描述符[25]。使用20多种不同生物中的约345 000个已识别肽来训练网络，经过训练得出了比优化前更好的1 303个肽的预测准确度。该算法的主要限制因素在于需要大量的训练肽，这使得其难以适用于其他色谱条件。

1.4.2　支持向量回归 (SVR)

为了达到使用较少的训练肽的同时也能适应不同的色谱条件，Moruz等[5]开发了一个基于SVR的RT预测算法Elude。Elude参数化了约60个氨基酸特征：氨基酸组成、肽长度、末端残基类型、高度带电的氨基酸残基、最近邻效应、疏水性（平均疏水性，N和C末端疏水性，最多或最少疏水性氨基酸的出现次数）、二级结构等。方法主要特点在于：在有足够训练肽数据的情况下，Elude直接构建一组线性保留指数，计算肽特征并使用SVR进行最佳组合，从而达到预测保留时间的目的。如果没有足够数据，Elude先运行少量对照肽，再从库中选择最合适（预测RT和观察RT的相关性最高）的模型并将其校准。使用对异常值处理比Pearson相关系数更稳健的FAST-最小修整平方（FAST-least trimmed squares，FAST-LTS）回归方法进行选择和校准。这种方案确保了该算法可以应用于不同的色谱条件，并保证了最小性能损失。

在此基础上又衍生出来许多SVR组合算法预测模型。串并行支持向量机（serial and parallel

support vector machine,SP-SVM)包含一个仅用于模型训练的SVR (p-SVR)和4个用于RT预测的SVM (C-SVM、1-SVR、s-SVR和n-SVR)[26]。其中,C-SVM计算肽色谱行为特征,1-SVR和s-SVR进行目标肽段RT预测,n-SVR对肽RT归一化,以表征多肽之间的相互作用,进一步提高了其预测准确度和性能。

不确定性可以公式化为目标样本与训练数据集之间的关系,所以掌握了这样的预测策略之后,GPTime便将SVR替代为高斯计算过程(Gaussian Processes,GP),以同样的选择－训练－校准－计算模式,证明了GP与SVR同等的准确性,同时提供了预测RT的不确定性估计[27]。

Lu等[28]从新的角度出发,提出了一个基因座特异性保留预测因子(locus-specific retention predictor,LsRP),它新颖地将氨基酸基因座信息与SVR算法结合。将每个肽序列转化为由0和1组成的特征基因座载体,使基因座载体和肽序列之间保持一对一的对应关系,再进行SVR训练和评估。LsRP最终提供了0.95 ~ 0.99的预测相关系数。

1.4.3 深度学习 深度学习可以自动从庞大数据中有效解读复杂关系并学习特征和模式,无需进行人工特征设计,因此特别适合大型的复杂数据集的科学领域。基于深度学习的算法大致分为3类:递归神经网络(recurrent neural network,RNN)、卷积神经网络(convolutional neural networks,CNN)和混合网络,其中RNN是最主要的网络架构。

Prosit是RNN的代表性算法[29],由一个编码器和一个解码器组成。编码器将肽序列编码为离散整数向量(每个氨基酸残基长度为20)的表示形式,而解码器则对该表示形式进行解码,预测RT。编码器由一个嵌入层,一个BiGRU层,一个递归GRU层和一个关注层组成[30]。解码器将序列的表示形式连接到密集层从而进行预测。同样基于RNN架构的DeepMass则使用一键编码,其网络包括一个BiLSTM层、一个LSTM层,两个致密层[31]。GuanMCP2019[32]则使用了一个屏蔽层、两个BiLSTM层、一个LSTM层,两个致密层。与SSRCalc和Elude比较,这几种算法都显示出优异的性能,对RT的预测可以达到接近1的相关性。

CNN包含卷积层和池化层,可在不同的空间尺度上提取序列特征。Ma等[33]提出DeepRT,是CNN和RNN的混合网络架构,其预测程序是:在特征自主学习(CNN层和LSTM层)之后,利用主成分分析(principal component analysis ,PCA)进行降维,然后利用3种常规机器学习方法(SVR,随机森林(random forest,RF),梯度提升(gradient boosting,GB))进行建模。Deep DIA[34]和Auto RT[35]都是这样的混合架构,区别是二者的RNN层分别为BiLSTM和GRU。值得一提的是,AutoRT有两个独特功能:其一是通过遗传算法实现自动深度神经网络体系架构搜索(network architecture search,NAS),从而识别出10个最匹配的模型进行组合预测;另一个就是转移学习,转移学习的特点是大型公共数据集的使用。使用大型公共数据集(约174 182条肽)对基础模型进行训练,然后用少量目标数据对基础模型进行微校准以适用于特定的实验条件。有这样的公共数据集在,即使实验数据量只有几百条也能够得到获得高度准确的模型。

对于较小的数据集,传统的机器学习方法通常优于深度学习方法,但是随着训练集的增多,深度学习方法的优势便逐渐显现,性能也大大优于机器学习[36]。

2 具有翻译后修饰多肽的保留时间预测

PTM能够改变蛋白质的电荷状态、疏水性、空间结构和稳定性,最终影响其与受体等的相互作用及功能。目前已发现300多种不同的PTM,主要形式包括磷酸化、糖基化、乙酰化、羧基化、糖基化以及二硫键的配对等[37]。PTM引起的肽RT的变化取决于修饰类型和数量,发生修饰的氨基酸残基种类及其在序列中的位置。

2.1 特定PTM修饰

目前有很多研究在开发适用于PTM肽的RT预测,大多是在已有模型基础上引入修饰的氨基酸残基的模型参数(RC,疏水性等)来进行预测。如Reimer[38]引入不同的几组修饰肽的保留数据,建立一种序列依赖性的方法来预测N端烷基化修饰的肽段。烷基化修饰使N末端残基的疏水性增加,表现出更强的保留。同时洗脱条件的变化对保留时间后移的影响也更为明显。

BioLCCC的拓展模型可以预测具有磷酸化修饰的肽[39]，天冬酰胺脱酰胺化修饰和天冬氨酸异构化修饰的肽[40]。当C_{18}柱与醋酸、甲酸（formic acid，FA）、或三氟乙酸（trifluoroacetate，TFA）等离子对试剂使用时，磷酸肽通常比它们的未磷酸化对应物表现出更强的保留，而当使用疏水性较小的固定相（如C4-硅胶柱）时，保留顺序逆转[41]。色谱条件的改变如RP C_{18}固定相的离子对试剂可能会影响其分离的选择性及预测准确性，用FA代替TFA则需要重新校准模型参数。未来的算法研究无疑需在流动相极性的影响上进行更多的探索。

2.2　任意PTM修饰

Elude 2.0[42]能够适用于任意PTM，前提是需要足够的数据来解释每种修饰氨基酸的特性。为了将其功能扩展到修饰肽，删除了疏水性指数Kyte-Doolittle，修改并添加了部分描述符，如25%最低和最高RC的发生次数/连续出现次数等。在RPLC-FA系统中，乙酰化、丁酰化和丙酰化修饰的肽通常在未修饰肽之后洗脱，甲硫氨酸、蛋氨酸氧化修饰的肽在未修饰肽之前洗脱。Elude2.0对修饰和未修饰的肽具有同样优异的预测性能，所有数据集的预测和实验RT之间的相关系数为0.93～0.98。由于未知肽段序列的每一个位点都可能存在修饰且会导致保留行为的差异，因此，为了准确扩展模型，需要在统计上大批量地、可靠地识别并数据化目标修饰肽段的RT。

在深度学习方法中，大多数模型采用的一键编码氨基酸的形式限制了PTM肽段的适用性[43]。DeepLC[44]是唯一可以预测修饰多肽RT的模型，甚至是训练集中不存在的修饰类型。DeepLC采用CNN架构，每种肽被编码为矩阵来计算其原子组成，对于含修饰氨基酸的多肽，修饰的原子组成直接加到未修饰残基的原子组成上。这种编码使模型能够学习并归纳未知的修饰肽段。考虑到异构体的存在，除此编码外，还编码了位置特定信息和全局特征信息，这使得Deep LC预测修饰肽段（尤其是酰基修饰）和未修饰肽段的RT准确度相当。在20个数据集中（SWATH Library29，HeLa HF30和DIA HF31等），Pearson相关系数都能达到0.99。但DeepLC对具有复杂修饰（磷酸化或异构化）的肽段进行RT预测还是较为困难，准确度较低，需要一些与复杂修饰相关的训练数据才能进一步提高性能。

3　应　用

在靶向蛋白质组学中，保留时间预测模型可以潜在地帮助生成数据采集的参考列表，实现更多的蛋白质同时定量。在bottom-up蛋白质组学中，这些模型主要用于在数据库搜索过程中，作为肽匹配图谱（peptide-spectrum matches，PSM）的额外验证标准。近年来，越来越多的研究将多肽RT预测模型集成到蛋白质组学数据分析工作流程中。这些不同原理的方法已大量应用于数据依赖采集（data dependent acquisition，DDA）靶向蛋白质组学实验、DIA蛋白质组学实验和完整蛋白质RT预测的综合模型开发中。

3.1　靶向蛋白质组学分析中的肽保留时间预测

对于靶向蛋白质组学中关键的第一步“方法开发建立”，预测的RT已用于减少分析靶标所需的实验次数。采集窗口越小，便可以在不损害数据质量的情况下靶向更多的肽。复杂的背景可能导致选择反应监测（selected reaction monitoring，SRM）测量结果的模糊性，因为样品中可能存在具有与目标肽段类似的干扰肽。在DDA中，Prosit[29]包含来自于576 256个母离子的21 764 501高质量谱图，覆盖98.5%的人类基因。使用预测得到的准确的RT和二级谱进行匹配打分，大大提高了对靶向肽段的检出能力（增加20%）。类似的还有基于SSRCalc的软件应用，简化了质谱仪方法的开发流程[45]，可测量酿酒酵母中MS可观察到的所有蛋白质（100%）[46]。随着色谱柱的变化或仪器中归一化碰撞能量（normalized collision energy，NCE）调谐漂移，基于DDA的谱库会随着时间的流逝而变得过时。

3.2　DIA蛋白质组学分析中的肽保留时间预测

二级谱是混合谱，DIA的数据来源于很多肽段，而且碎片离子还会受到未碎裂的母离子的干扰，在短色谱梯度与复杂样品同时出现的情况下，干扰会进一步被放大。在没有碎片谱图提供的高可信度数据的情况下，可以将观察到的肽段RT和未碎片化的质量用作肽段鉴定的附加信息，过滤错误识别的代谢产物。这些预测算法的优势在于可以确保库始终是最新的，甚至可以考虑不同仪器平台之间的差异。DIA方法思路大致为，使用相似

样品来源(如酿酒酵母蛋白质)数据库及Prosit14辅助生成RT预测的谱库(320 150个独特的肽序列),经过经验校正(6次气相分馏DIA进样),新库包含来自4 464个蛋白质组的64 597个肽序列[47]。肽和蛋白质的FDR为1%。每种肽采集后从库中选择得分最高的电荷状态,删除其他得分较低的电荷状态。然后,对于每个鉴定出的肽,计算所有碎片离子的总峰形,并提取与该形状相关的所有可能的b型或y型离子的碎片峰面积强度进行定量。

高精度iRT[12]能够实现在很多不同色谱系统下,将肽段保留时间特征转换成精确可预测的时间信息,从而高精度地预测肽段RT,实现更多蛋白质的同时定量(增加25%)。只需一次靶标肽段的校正实验,形成新的iRT计算器,就可计算该色谱系统下的待测肽段的保留时间预测值。随后便可利用得到的所有待测肽段的预测RT,设计适合的梯度靶向分析方法,提高更多峰鉴定的可靠性。Klammer等[24]基于SVR算法对酿酒酵母细胞裂解物的检出率增加了50%,FDR降低至3%。Moruz等[48]基于Elude算法分别在酵母和人类的两个三重数据集上进行了评估,在FDR为1%的情况下,蛋白质的鉴定检出率多出了7%。目前在DIA中,较有前景和应用空间的是深度学习算法,在此类模型中,可从经验示例中了解肽序列(或衍生自该序列的特征)与LC保留时间顶点之间的映射。Prosit[29]可以直接用于DIA的建库(FDR=1%)。DeepDIA[34]构建了计算机模拟血浆/血清蛋白质组库,平均检测到的蛋白质组超过400个,是从相同数据中基于最新DDA库检测到的蛋白质组的两倍。DIA-NN[49]可通过短色谱梯度实现可靠的鉴定和深度蛋白质组学覆盖。其基于深度神经网络进行量化和干扰校正,来区分真实信号和噪声。使用iRT进行保留时间校准,同时自动执行质量校正。在考虑0.5% FDR进行采集的情况下,比基于SSRCalc的方法识别出K562人细胞系全细胞胰蛋白酶消化物更多的前体肽段(约35 000个)。

3.3 完整蛋白的保留时间预测

丰富的多肽保留预测模型的经验能够应用在完整蛋白质的RT预测上,当然也更具挑战性。Bio LCCC,基于高分子统计物理学方法,把吸附剂孔内的所有多肽链分子的可能构型都考虑在内,对于完整蛋白质的RT预测有良好的可行性。研究表明,BioLCCC模型在12个完整蛋白质[50]和52个完整蛋白质[51](氨基酸残基数多达583)的数据集中,实验RT和预测RT之间的相关性可达到0.89和0.90。但该方法的不足之处就在于其针对的是链状结构,对于含二级三级结构的蛋白质来说,相关性有待进一步提高。不局限于RPLC,Xu等[52]使用偏最小二乘回归将模型蛋白质的等电点,相对分子质量和水两相分配系数与阳离子交换色谱(ion-exchange chromatography,IEC)的RT相关联。对9种蛋白质进行训练时获得0.91的线性相关性。此外,疏水相互作用色谱(hydrophobic interaction chromatography,HIC)是蛋白质分离纯化的关键技术。Chen等[53]生成了基于SVM方法的定量结构特性关系(quantitative structure property relationship,QSPR)模型,使用氨基酸组成来估算有效的蛋白质疏水性,用于预测模型中未包含的蛋白质的等度及进一步的线性梯度保留参数。对于20个蛋白质的数据集,实验RT和预测RT之间的相关性可达到0.97。而定量结构活性关系(quantitative structure activity relationship,QSAR)则使用同源性建模和分子动力学模拟来生成单克隆抗体(monoclonal antibodies,mAbs)的3D结构,然后从中计算结构描述符以预测mAbs的HIC保留时间[54]。

4 总结与展望

在基于LC-MS技术的蛋白质组学中,保留时间对多肽鉴定及定量的准确性、完整性和深入性起到重要作用。与基于多肽分子模型的方法相比,索引及序列特异性模型的应用性更广泛,但其预测能力仍受限于色谱条件。随着研究的不断深入,在更多数据集、更多未知肽段及蛋白面前,通过训练深度神经网络模型,构建专属于每一台仪器的网络模型或组合模型,采集时间可以从几天大大缩短至几小时。PTM修饰肽的保留模型的发展未来集中在,在训练集中无已知修饰类型的参数的前提下,优化由空间结构变化导致的修饰这一方面的数据。在多肽RT预测领域,仍需进一步提高模型的准确性,建立统一的评价标准,开发更具普适性的算法,使RT预测真正成为蛋白质组学研究的重要手段之一。

References

[1] Henneman A, Palmblad M. Retention time prediction and protein identification[J]. *Methods Mol Biol*, 2020, **2051**: 115-132.

[2] Dorfer V, Maltsev S, Winkler S, *et al*. CharmeRT: boosting peptide identifications by chimeric spectra identification and retention time prediction[J]. *J Proteome Res*, 2018, **17**(8): 2581-2589.

[3] Escher C, Reiter L, Maclean B, *et al*. Using iRT, a normalized retention time for more targeted measurement of peptides[J]. *Proteomics*, 2012, **12**(8): 1111-1121.

[4] Krokhin O, Craig R, Spicer V, *et al*. An improved model for prediction of retention times of tryptic peptides in ion pair reversed-phase HPLC: its application to protein peptide mapping by off-line HPLC-MALDI MS [J]. *Mol Cell Proteomics*, 2004, **3**(9): 908-919.

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

[6] Zohora FT, Rahman MZ, Tran NH, *et al*. DeepIso: a deep learning model for peptide feature detection from LC-MS map[J].*Sci Rep*, 2019, **9**(1): 17168.

[7] Baczek T, Kaliszan R, Novotná K, *et al*. Comparative characteristics of HPLC columns based on quantitative structure-retention relationships (QSRR) and hydrophobic-subtraction model [J]. *J Chromatogr A*, 2005, **1075**: 109-115.

[8] Le Maux S, Nongonierma A, Fitzgerald R. Improved short peptide identification using HILIC-MS/MS: retention time prediction model based on the impact of amino acid position in the peptide sequence[J]. *Food Chem*, 2015, **173**: 847-854.

[9] Gorshkov A, Tarasova I, Evreinov V, *et al*. Liquid chromatography at critical conditions: comprehensive approach to sequence-dependent retention time prediction[J]. *Anal Chem*, 2006, **78**(22): 7770-7777.

[10] Tarasova I, Goloborodko A, Perlova T, *et al*. Application of statistical thermodynamics to predict the adsorption properties of polypeptides in reversed-phase HPLC[J]. *Anal Chem*, 2015, **87**(13): 6562-6569.

[11] Gallien S, Peterman S, Kiyonami R, *et al*. Highly multiplexed targeted proteomics using precise control of peptide retention time[J]. *Proteomics*, 2012, **12**(8): 1122-1133.

[12] Bruderer R, Bernhardt O, Gandhi T, *et al*. High-precision iRT prediction in the targeted analysis of data-independent acquisition and its impact on identification and quantitation[J]. *Proteomics*, 2016, **16**: 2246-2256.

[13] Meek J. Prediction of peptide retention times in high-pressure liquid chromatography on the basis of amino acid composition [J]. *Proc Natl Acad Sci U S A*, 1980, **77**(3): 1632-1636.

[14] Mant C, Hodges R. Context-dependent effects on the hydrophilicity/hydrophobicity of side-chains during reversed-phase high-performance liquid chromatography: implications for prediction of peptide retention behaviour[J]. *J Chromatogr A*, 2006, **1125**(2): 211-219.

[15] Mant C, Hodges R. Design of peptide standards with the same composition and minimal sequence variation to monitor performance/selectivity of reversed-phase matrices[J]. *J Chromatogr A*, 2012, **1230**: 30-40.

[16] Tripet B, Cepeniene D, Kovacs JM, *et al*. Requirements for prediction of peptide retention time in reversed-phase high-performance liquid chromatography: hydrophilicity/hydrophobicity of side-chains at the N- and C-termini of peptides are dramatically affected by the end-groups and location[J]. *J Chromatogr A*, 2007, **1141**(2): 212-225.

[17] Dwivedi R, Spicer V, Harder M, *et al*. Practical implementation of 2D HPLC scheme with accurate peptide retention prediction in both dimensions for high-throughput bottom-up proteomics [J]. *Anal Chem*, 2008, **80**(18): 7036-7042.

[18] Reimer J, Spicer V, Krokhin O. Application of modern reversed-phase peptide retention prediction algorithms to the Houghten and DeGraw dataset: peptide helicity and its effect on prediction accuracy[J]. *J Chromatogr A*, 2012, **1256**: 160-168.

[19] Spicer V, Lao Y, Shamshurin D, *et al*. N-capping motifs promote interaction of amphipathic helical peptides with hydrophobic surfaces and drastically alter hydrophobicity values of individual amino acids[J]. *Anal Chem*, 2014, **86**(23): 11498-11502.

[20] Krokhin O, Ezzati P, Spicer V. Peptide retention time prediction in hydrophilic interaction liquid chromatography: data collection methods and features of additive and sequence-specific models[J]. *Anal Chem*, 2017, **89**(10): 5526-5533.

[21] Gussakovsky D, Neustaeter H, Spicer V, *et al*. Sequence-specific model for peptide retention time prediction in strong cation exchange chromatography [J]. *Anal Chem*, 2017, **89**(21): 11795-11802.

[22] Petritis K, Kangas L, Ferguson P, *et al*. Use of artificial neural networks for the accurate prediction of peptide liquid chromatography elution times in proteome analyses[J]. *Anal Chem*, 2003, **75**(5): 1039-1048.

[23] Shinoda K, Sugimoto M, Yachie N, *et al*. Prediction of liquid chromatographic retention times of peptides generated by protease digestion of the *Escherichia coli* proteome using artificial neural networks[J]. *J Proteome Res*, 2006, **5**(12): 3312-3317.

[24] Klammer AA, Yi X, Maccoss MJ, *et al*. Peptide retention time prediction yields improved tandem mass spectrum identification for diverse chromatography conditions [J]. *Anal Chem*, 2007, **79**(16): 6111-6118.

[25] Petritis K, Kangas L, Yan B, *et al*. Improved peptide elution time prediction for reversed-phase liquid chromatography-MS by incorporating peptide sequence information[J]. *Anal Chem*, 2006, **78**(14): 5026-5039.

[26] Zhang J, Zhang D, Zhang W, *et al*. A new peptide retention time

prediction method for mass spectrometry based proteomic analysis by a serial and parallel support vector machine model[J]. *Se Pu*, 2012, **30**(9): 857-863.

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

[28] Lu W, Liu X, Liu S, *et al*. Locus-specific retention predictor (LsRP): a peptide retention time predictor developed for precision proteomics[J]. *Sci Rep*, 2017, **7**: 43959.

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

[30] Fergadis A, Baziotis C, Pappas D, *et al*. Hierarchical bi-directional attention-based RNNs for supporting document classification on protein-protein interactions affected by genetic mutations.[J] .*Database* (*Oxford*), 2018: bay076.

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

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

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

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

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

[36] Bouwmeester R, Gabriels R, Van Den Bossche T, *et al*. The age of data-driven proteomics: how machine learning enables novel workflows[J]. *Proteomics*, 2020, **20**(21/22): e1900351.

[37] Olsen J, Mann M. Status of large-scale analysis of post-translational modifications by mass spectrometry [J]. *Mol Cell Proteomics*, 2013, **12**(12): 3444-3452.

[38] Reimer J, Shamshurin D, Harder M, *et al*. Effect of cyclization of N-terminal glutamine and carbamidomethyl-cysteine (residues) on the chromatographic behavior of peptides in reversed-phase chromatography[J]. *J Chromatogr A*, 2011, **1218**(31): 5101-5107.

[39] Perlova T, Goloborodko A, Margolin Y, *et al*. Retention time prediction using the model of liquid chromatography of biomacromolecules at critical conditions in LC-MS phosphopeptide analysis[J]. *Proteomics*, 2010, **10**(19): 3458-3468.

[40] Sargaeva NP, Goloborodko AA, O′connor PB, *et al*. Sequence-specific predictive chromatography to assist mass spectrometric analysis of asparagine deamidation and aspartate isomerization in peptides[J]. *Electrophoresis*, 2011, **32**(15): 1962-1969.

[41] Ogata K, Krokhin O, Ishihama Y. Retention order reversal of phosphorylated and unphosphorylated peptides in reversed-phase LC/MS[J]. *Anal Sci*, 2018, **34**(9): 1037-1041.

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

[43] Wen B, Zeng W, Liao Y, *et al*. Deep learning in proteomics[J]. *Proteomics*, 2020, **20**(20/21): e1900335.

[44] Ivanov MV, Bubis JA, Gorshkov V, *et al*. Boosting MS1-only proteomics with machine learning allows 2000 protein identifications in single-shot human proteome analysis using 5 min HPLC gradient[J] .*J Proteome Res*, 2021, **20**(4): 1864-1873.

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

[46] Röst H, Malmström L, Aebersold R, *et al*. A computational tool to detect and avoid redundancy in selected reaction monitoring [J]. *Mol Cell Proteomics*, 2012, **11**(8): 540-549.

[47] Searle BC, Swearingen KE, Barnes CA, *et al*. Generating high quality libraries for DIA MS with empirically corrected peptide predictions[J] .*Nat Commun*, 2020, **11**(1): 1548.

[48] Moruz L, Hoopmann M, Rosenlund M, *et al*. Mass fingerprinting of complex mixtures: protein inference from high-resolution peptide masses and predicted retention times [J]. *J Proteome Res*, 2013, **12**(12): 5730-5741.

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

[50] Gorshkov AV, Evreinov VV, Pridatchenko ML, *et al*. Applicability of the critical-chromatography concept to analysis of proteins: dependence of retention times on the sequence of amino acid residues in a chain[J]. *Polymer Sci*, 2011, **53**(12): 1227-1241.

[51] Pridatchenko M, Perlova T, Ben Hamidane H, *et al*. On the utility of predictive chromatography to complement mass spectrometry based intact protein identification[J]. *Anal Bioanal Chem*, 2012, **402**(8): 2521-2529.

[52] Xu L, Glatz C. Predicting protein retention time in ion-exchange chromatography based on three-dimensional protein characterization[J]. *J Chromatogr A*, 2009, **1216**(2): 274-280.

[53] Chen J, Yang T, Cramer S. Prediction of protein retention times in gradient hydrophobic interaction chromatographic systems [J]. *J Chromatogr A*, 2008, **1177**(2): 207-214.

[54] Karlberg M, de Souza JV, Fan L, *et al*. QSAR Implementation for HIC retention time prediction of mAbs using fab structure: a comparison between structural representations [J] .*Int J Mol Sci*, 2020, **21**(21): 8037.