2010 年 6 月 色 谱 Vol. 28 No. 6
June 2010 Chinese Journal of Chromatography 529 ~ 534

专论与综述

DOI:10.3724/SP.J.1123.2010.00529

肽段反相色谱保留时间预测算法及其在蛋白质鉴定中的应用

刘 超[1,2], 王海鹏[1,2], 付 岩[1], 袁作飞[1,2], 迟 浩[1,2],
王乐珩[1], 孙瑞祥[1*], 贺思敏[1*]
(1. 中国科学院计算技术研究所智能信息处理重点实验室, 北京 100190;
2. 中国科学院研究生院, 北京 100049)

摘要:液相色谱-质谱(LC-MS)联用是当今规模化蛋白质鉴定的主流技术。肽段在反相液相色谱(RPLC)中的保留时间主要是由肽段的理化性质和 LC 条件(固定相、流动相)决定的。可以通过分析肽段的理化性质,并量化它们对肽段色谱行为的影响来预测保留时间。预测结果可以用于帮助提高蛋白质鉴定的数量和可信度,也可用于肽段的翻译后修饰等研究。现在已有的保留时间预测算法主要有保留系数法和机器学习法两大类,得到的预测保留时间与实际保留时间相关系数可达到 0.93。随着色谱和质谱技术的不断发展,肽段色谱行为的稳定性和重现性越来越好,保留时间预测结果也越来越准确。预测肽段保留时间将成为提高蛋白质鉴定结果的重要技术手段之一。

关键词:反相液相色谱-质谱联用;保留系数;机器学习;保留时间;预测;蛋白质;肽;鉴定
中图分类号:O658 **文献标识码**:A **文章编号**:1000-8713(2010)06-0529-06

Prediction of peptide retention time in reversed-phase liquid chromatography and its application in protein identification

LIU Chao[1,2], WANG Haipeng[1,2], FU Yan[1], YUAN Zuofei[1,2], CHI Hao[1,2],
WANG Leheng[1], SUN Ruixiang[1*], HE Simin[1*]
(*1. Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China; 2. Graduate University of Chinese Academy of Sciences, Beijing 100049, China*)

Abstract: Liquid chromatography-mass spectrometry (LC-MS) is the mainstream of high-throughput protein identification technology. Peptide retention time in reversed-phase liquid chromatography (RPLC) is mainly determined by the physicochemical properties of the peptide and the LC conditions (stationary phase and mobile phase). Retention time can be predicted by analyzing these properties and quantifying their effects on peptide chromatographic behavior. Prediction of peptide retention time in LC can be used to improve identification of peptides and post translational modifications (PTM). There are mainly two methods to predict retention time: i.e. retention coefficients and machine learning. The coefficient of determination between observed and predicted retention times can reach 0.93. With the development of LC-MS technology, retention time prediction will become an important tool to facilitate protein identification.

Key words: reversed-phase liquid chromatography-mass spectrometry (RPLC-MS); retention coefficient; machine learning; retention time; prediction; protein; peptide; identification

高效液相色谱-质谱(HPLC-MS)联用作为蛋白质鉴定的重要技术,在蛋白质组学中得到了广泛应用[1]。但由于蛋白质组的复杂性和基于质谱技术的蛋白质鉴定方法的局限性,鉴定结果的数量和可

* 通讯联系人:孙瑞祥,博士,副研究员,主要研究方向为计算蛋白质组学. Tel:(010)62601016, E-mail:rxsun@jdl.ac.cn.
贺思敏,博士,研究员,主要研究方向为计算蛋白质组学. Tel:(010)62601016, E-mail:smhe@jdl.ac.cn.
基金项目:国家"863"计划项目(No. 2008AA02Z309 和 No. 2007AA02Z315)、国家"973"计划项目(No. 2010CB912701)和中国科学院知识创新计划项目(No. KGGX1-YW-13).
收稿日期:2010-01-26

信度一直无法让人完全满意[2]。现在的一些方法，如谱图聚类[3]、多引擎归并[4]、谱图预处理[5]等，主要是利用从一级和串联谱图中提取的信息来提高鉴定结果的数量和可信度。最近的研究表明，从一维反相液相色谱（RPLC）中得到的肽段保留时间信息，也可用于提高肽段鉴定数量和可信度[6]。

保留时间是指组分从进入液相色谱柱到出现最大峰值所需的时间，它是独立于 MS 信息以外的另一维重要信息。在一定的 RPLC 条件下，不同的肽段其保留时间不同。根据肽段的序列等信息，可对肽段在色谱中的保留时间进行预测。现在的预测算法得到的保留时间与实际保留时间的相关系数（R^2）已经达到 0.93[6]。

对于通过质谱手段[7]（包括数据库搜索，从头测序和序列标签等）鉴定出的肽段，可以采用保留时间预测算法得到它们的预测保留时间，如果预测的保留时间与实际保留时间一致（误差小于某个阈值），那么该鉴定结果的可信度就较高；如果相差很大，那么该结果可能是假阳性的，需进一步地确认。

发生在不同修饰位点的不同修饰基团对肽段保留时间的影响各异，通过分析、预测这些差异，可以更好地确定肽段发生的翻译后修饰（post-translational modification，PTM）类型和位点。

本文主要介绍常见的肽段 RPLC 保留时间预测算法及其在蛋白质鉴定中的应用，并对建立预测算法需要注意的细节进行了讨论。

1 保留时间预测方法

1.1 问题描述

肽段在一定色谱条件中的保留时间是由肽段的理化性质决定的，如氨基酸疏水性、肽段长度、肽段分子质量等。如果这些理化性质对保留时间的影响是稳定且可量化的，那么就可以构造预测函数 $t_{\mathrm{pre}} = f(p,\theta)$ 来预测保留时间，其中 p 表示一个肽段，θ 表示函数参数，t_{pre} 是预测保留时间。

早期的研究使用人工合成肽段来分析和预测保留时间。随着蛋白质组学的发展，在现在常见的鸟枪法（shotgun）蛋白质质谱鉴定实验中，可以使用鉴定出的肽段 $\{p_i \mid i = 1,2,\cdots,n\}$ 以及它们的实际保留时间 $\{t_i \mid i = 1,2,\cdots,n\}$ 去训练预测函数 $t_{\mathrm{pre}} = f(p,\theta)$，并得到对应的预测保留时间 $\{t_i' \mid i = 1,2,\cdots,n\}$。通常使用预测保留时间 $\{t_i' \mid i = 1,2,\cdots,n\}$ 与实际保留时间 $\{t_i \mid i = 1,2,\cdots,n\}$ 的相关系数 R^2[8] 来评价预测算法的有效性，相关系数越高，算法的预测效果越好。

预测保留时间的关键是构建预测算法，包括选择预测函数和确定相关参数。

1.2 基于氨基酸保留系数的方法

1980 年，Meek[9] 提出了保留系数（retention coefficients，RC）方法，通过将组成肽段的氨基酸保留系数加和来预测该肽段的保留时间，如下式：$T = \sum R_i N_i + b$，其中 T 是该肽段的保留时间，R_i 是第 i 种氨基酸（本文提到的肽段中氨基酸一般都指氨基酸残基）的保留系数；N_i 是第 i 种氨基酸在该肽段中的数目，生物体内的常见氨基酸有 20 种，故 $i = 1,2,\cdots,20$。在一定的色谱条件下，b 是常数，表示色谱死时间或系统偏差等。这里 R_i 和 b 对应 1.1 节中提到的预测函数中的输入参数 θ；N_i 是肽段自身的属性，对应预测函数中的输入参数 p。

在常见的 RPLC 实验中，疏水性越强的氨基酸保留系数的值越大，表示含有这类氨基酸的肽段与色谱固定相结合性越强，需要经过较长时间、移动相有机溶剂浓度增大到一定值时才能被洗脱下来；而亲水性氨基酸保留系数小，表示含有这类氨基酸的肽段容易被洗脱，保留时间短。在该算法中，确定氨基酸的保留系数是关键。Meek[9] 使用 25 个肽段进行启发式计算，得到了 20 种氨基酸的保留系数，其中保留系数最大的为苯丙氨酸（13.4），最小的为丝氨酸（－3.2）。Browne 等[10] 使用类似的算法，在不同的色谱条件下重新计算了各氨基酸的保留系数。其实验表明，不同的色谱条件下，氨基酸的保留系数不尽相同，需要进行专门计算，区别对待。

Guo 等[11,12] 使用合成的肽段，考虑不同的色谱条件（填料、色谱柱长度、pH 值和温度等），对 RC 法进行进一步的实验与分析，建立了一套比较系统的预测算法。该方法中使用的合成肽段形式为 Ac-GXXLLLKK-NH_2，其中 X 表示 20 种氨基酸中的一种，如 Ac-GAALLLKK-NH_2、Ac-GCCLLLKK-NH_2。在每次实验中，X 被替换为某种氨基酸，根据各合成肽段的保留时间差异，计算氨基酸的保留系数。虽然该方法的预测结果并不十分理想（$R^2 = 0.81$），但为以后发展基于复杂蛋白质混合样品的 RC 算法奠定了基础。

Casal 等[13] 使用分段最小二乘法（partial least squares，PLS）和多元线性回归法（multiple linear regression，MLR）确定保留系数，预测效果更好，证明回归分析方法在预测保留时间方面更有效。

2002 年，Palmblad 等[14] 将预测保留时间引入蛋白质组学的研究，尝试结合保留时间预测来提高肽指纹图谱法（peptide mass fingerprinting，PMF）

的蛋白质鉴定结果。他们的样品为牛血清蛋白、人血清和志愿者提供的人脑脊液,预测方法为基于最小二乘(LS)的 RC 法。该实验表明复杂生物样品酶切出的肽段的保留时间也是可预测的。Palmblad 等[15]后来又使用了更复杂的样品(酶切出的肽段更多),对预测保留时间在蛋白质鉴定中的应用进行了更加细致的分析。人脑脊液实验结果表明,在相同假阳性结果数量下,通过预测保留时间,鉴定到的蛋白质数量(真阳性结果)提高了一倍。

RC 法认为肽段的保留时间和氨基酸的保留系数为线性关系,因此可以使用线性回归等方法,如 LS 法来预测保留时间。虽然使用 RC 法得到的预测效果不错,但是保留时间与各氨基酸属性未必是线性关系,RC 法是对未知函数 $t_{pre}=f(p,\theta)$ 的简化,缺乏理论基础。另外,使用这种方法得到的同分异构肽段的保留时间是相同的,而实际上,同分异构肽段的保留时间一般是不同的。

1.3　基于氨基酸保留系数及肽段理化性质的方法

除了各氨基酸自身的理化性质,肽段的长度、相对分子质量、各氨基酸间的关系也会对肽段保留时间有影响。Mant 等[16]分析了肽段长度对保留时间的影响,认为在使用 RC 法预测保留时间时,不同长度的肽段要区别对待。Mant 和 Hodges[17]最近的工作表明,氨基酸 K 和 R 的侧链的疏水性会随着整个肽段的疏水性的增加而增加;肽段所带正电荷数越多,氨基酸 I 的侧链的疏水性越弱。Tripet 等[18]对肽段 N 端和 C 端氨基酸侧链进行了研究,通过每次替换肽段中一种氨基酸的方法,得到了氨基酸在 N 端和 C 端的保留系数,并与氨基酸在肽段中除 N、C 端外的其他(中间)位置时的保留系数(internal coefficients)对比,发现氨基酸在 C 端时,保留系数变化较大,而在 N 端时只有微小的变化。

Houghten 和 DeGraw[19]对氨基酸在肽段中的相对位置进行了研究,发现氨基酸组成相同而位置不同的肽段,保留时间并不一致。Zhou 等[20]观察到两亲性 α-螺旋(amphipathic α-helical)结构对肽段的保留时间有较大影响,与氨基酸一样,也应该对该结构进行保留系数预测。

2004 年,Krokhin 等[21]在 Guo 等[11,12]工作的基础上,考虑了肽段长度、疏水性以及氨基酸在肽链上的分布等信息对肽段保留时间的影响因素,基于 17 个蛋白质(共 346 个酶解肽)混合物产生的数据集,根据经验值人工优化算法参数,提出了改进的保留时间预测算法 SSRCalc(sequence-specific retention calculator, http://hs2.proteome.ca/SSRCalc/SSRCalc.html)。该方法先用 RC 法得到一个初步的预测结果,然后根据色谱条件和肽段的其他理化性质对该预测结果进行进一步的校正和处理,得到最后的预测保留时间。SSRCalc 算法集成了前人的研究成果,使用的保留系数等主要参数都经过人工优化,考虑了较多的肽段保留时间的影响因素,是现在实用性最好的保留时间预测算法之一。与氨基酸保留系数一样,以上提到的各因素也是预测函数 $t_{pre}=f(p,\theta)$ 中的输入参数,而且它们与保留时间也未必是线性关系。SSRCalc 算法实质上是考虑不同影响因素的分段线性算法。要想进一步提高预测准确度,除了考察影响肽段保留时间的因素,也需要对预测函数的形式进行更加深入的研究。

Krokhin[22]使用更大的数据集(大约 2 000 个肽段),考虑更多的因素(最近邻效应、等电点和短序列肽段等),对该算法进行了改进,特别比较了不同色谱条件下算法参数的差异,使预测效果达到 $R^2=0.98$。Krokhin 等[23]还应用预测保留时间提高蛋白质鉴定的数量和可信度。在这些研究的基础上,2008 年,Dwivedi 等[24]提出了一种新的二维色谱离线分离方法。该方法先在碱性条件下对样品进行第一维 RPLC 分离,然后再在酸性条件下进行第二维 RPLC 分离。与常见的二维离子交换-反相色谱(SCX-RPLC)分离方法相比,这种方法除了具有分离效果好、实验和维护成本低的优点外,它的两维分离过程都可以进行保留时间预测,这样可结合以前发展的方法[23],提高鉴定结果的数量和可信度。在一次 80 h 的串联四极杆飞行时间质谱仪质谱实验中,一共可鉴定到大约 3 000 个可信蛋白质。

1.4　机器学习法

近几年,随着机器学习(machine learning, ML)理论的发展,人们开始使用 ML 方法进行保留时间预测[25]。虽然在 RC 法中,可以应用 ML 等复杂方法确定保留系数,但预测函数依然是线性形式。本节介绍的 ML 法是将各氨基酸的保留系数作为预测函数输入参数的一部分,并且又考虑了其他因素,得到的预测函数为非线性形式,预测效果也有一定程度的提高。

Petritis 等[26]使用人工神经网络(artificial neural network, ANN)对保留时间进行预测,该方法将组成肽的 20 种氨基酸数量作为输入值,输出该肽段的预测保留时间,使用 7 000 个已鉴定的高可信度肽段作为数据集进行交叉验证,得出每种氨基酸在算法中所对应的权值。2006 年,Petritis 等[27]对上述算法进行了改进,先后考虑了肽段长度、序

列、疏水性、疏水力矩、最近邻氨基酸和肽段二级结构等因素，使用从 12 059 种组织中鉴定出的约 345 000 个肽段进行训练，最后得到包含 1 052 个输入节点、24 个隐藏节点和 1 个输出节点的更精确的 ANN 保留时间预测算法。该算法能得出比较精确的预测保留时间，但训练时需要规模庞大的数据集。Shinoda 等[28]对氨基酸对 LC 保留时间的影响因素进行了统计分析和选择，并结合 ANN 和逐步多元线性回归（stepwise multiple linear regressions，SMLR）方法来预测保留时间。该方法可以准确预测由 50 个氨基酸组成的肽段的保留时间，对一个由 834 个肽段组成的数据集的预测结果为 $R^2 = 0.928$。

2007 年，Klammer 等[29]使用支持向量回归（support vector regressor，SVR）方法建立了保留时间预测算法。该算法使用含有 63 个元素的向量作为输入参数，其中 20 个元素分别表示 20 种氨基酸在肽段中的数量，40 个二值元素用来标识 N 端（N-terminal）和 C 端（C-terminal）是哪两种氨基酸，1 个元素表示 C 端最顶端的氨基酸（对于胰岛素酶切样品为 K 或 R），1 个元素表示肽段长度，最后一个元素表示肽段的相对分子质量。算法输出为预测保留时间。

ML 法的普遍适用性使得该方法也可以用于预测肽段在强阳离子交换（strong anion-exchange，SAX）色谱中的行为。因为本文主要讨论肽段在 RPLC 中的保留时间预测算法，这里不再展开讨论，可参见文献[30,31]。

ML 法中的 $t_{pre} = f(p, \theta)$ 不再是简单的线性函数，因而比 RC 法更符合实际，也能使用更多的氨基酸属性。但这类算法设计复杂，需要较大规模的训练集和测试集。RC 法的相关研究为发展 ML 法奠定了基础，而 ML 法吸收了 RC 法的成果，扩大了适用范围，取得了更为准确的预测结果。表 1 列出了几个比较有代表性的算法所考虑的影响因素。

表 1 对比不同算法所考虑的肽段保留时间的主要影响因素

Table 1 Comparison of effect factors of peptide retention time considered in different algorithms

Reference	Retention coefficient	Length[1)]	Mass	N-terminal[2)]	C-terminal[3)]	Sequence-dependent effect[4)]	Secondary structure[5)]
[9]	√						
[11,12]	√	√	√				
[21]	√	√		√		√	
[27]	√	√		√	√	√	√
[29]	√	√	√	√	√		

1) the number of amino acids of the peptide; 2) the retention coefficient of the amino acid at the N-terminal of the peptide; 3) the retention coefficient of the amino acid at the C-terminal of the peptide; 4) amphipathicity and nearest neighbor; 5) α-helix, β-sheet and coil.

2 保留时间预测在蛋白质鉴定中的应用

2.1 鉴定结果验证

在鸟枪法蛋白质质谱鉴定实验中，蛋白质鉴定的结果可以用于预测保留时间，而保留时间预测反过来可以校正蛋白质鉴定结果。

对于数据库搜索软件（如 SEQUEST[32]，MASCOT[33]，pFind[34]）给出的鉴定结果，设定保留时间误差阈值 ξ，当鉴定到的一个肽段的预测保留时间与实际保留时间之差（绝对值）大于 ξ 时，可认为该结果是假阳性的，反之是真阳性的，这样更好地区分了假阳性与真阳性鉴定结果，从而降低了假阳性率。为了应用保留时间预测方法，Krokhin 等[23]设计了一种迭代式的蛋白质鉴定流程。该流程首先从一级质谱中挑选强度最高的 5 个峰，并分析、鉴定它们对应的二级谱图，得到对应的肽段。对该肽段进行保留时间预测，若实际值与预测值相差小于 2 min，并且实验与理论母离子质量差小于 15 ppm，则认为鉴定出的肽段是可信的，将其添加到最终的鉴定结果中；否则是不可信的，将被排除掉。反复在一级质谱剩下的离子峰中选择 5 个强度最高的离子峰，重复上次过程，直到再也无法鉴定到新的蛋白质。在一次实验中，虽然通过这种方法实际采集到的二级谱数下降了 57%，但是鉴定结果的可信度更高，并且鉴定到的蛋白质数量并未减少。

在 Petrits 等[26]工作的基础上，Strittmatter 等[35]发展了用于过滤 SEQUEST[32]鉴定结果的判别函数（discriminate function），肽段的实际保留时间与预测保留时间的差异是其中的一个重要参数。该工作表明，实验与预测保留时间的差异可以有效地用于提高鉴定结果的数量。在人血清蛋白的实验中，可信鉴定结果的数量可提高 16%。

Klammer 等[29]从鉴定结果中提取比较可信的结果进行算法训练和阈值检验，然后过滤其他鉴定结果，从而降低了假阳性率，提高了鉴定的准确性。挑选可信鉴定结果（当数据数量有限时，可使用交叉验证的方法）进行算法训练和阈值检验，然后过滤更多的鉴定结果，是现在应用预测保留时间提高

蛋白质鉴定结果阳性率的普遍思路。除了预测保留时间,Sun 等[36]提出建立肽段保留时间库(empirical peptide retention time database)来过滤鉴定结果。将以前实验中的高可信肽段处理后加入到数据库中,新的实验结果与数据库中的比对,保留时间相差大的认为是假阳性结果。使用此方法过滤后,可信肽段数目最多提高了 60.8%。

2.2　PTM 研究

PTM 是蛋白质组学领域的重要研究方向,是蛋白质分子生物合成和功能实现的重要步骤。对于发生 PTM 的肽段,修饰基团会对肽段的疏水性产生影响,并导致该肽段的保留时间与相同序列的未发生修饰的肽段不同。如发生氧化修饰的肽段通常会比没发生氧化修饰的保留时间要小[37],而脱酰氨基修饰的肽段保留时间会稍许增加[38]。Krokhin 等[21]报道,在相同色谱条件下,糖基化的肽段比没有糖基化的肽段的保留时间少 2 min。

Kim 等[39]研究了肽段发生磷酸化后的保留时间偏移,发现一个磷酸化位点产生的偏移在 -5.28 min 和 +0.59 min 之间。他们还比较了不同磷酸化位点、不同磷酸化位点数目和不同色谱离子对试剂对肽段保留时间的影响。对于较短(氨基酸个数小于 18)的肽段,发生磷酸化的与未发生磷酸化的保留时间偏差 $\Delta t = (0.010\,4N - 0.176\,1)t_0$,其中 N 是肽段长度,t_0 是未发生磷酸化肽段的保留时间。

乙酰基和氨基官能团对肽段的保留时间也有影响。Baczek 等[40]研究了这种影响,发现肽段发生乙酰化后保留时间会增加,氨基官能团则会缩短肽段的保留时间。

预测发生 PTM 的肽段的保留时间,可先预测该肽段被修饰之前的保留时间,然后考虑修饰对保留时间的影响,对预测结果进行校正,校正后的结果即为修饰肽段的预测保留时间。Baczek 等[41]的预测算法考虑了肽段的结构信息,并将 PTM 作为结构信息的一部分,从而对修饰肽段进行保留时间预测。

2.3　应用时需要考虑的问题

被普遍认可的肽段保留时间影响因素有:a. 组成肽段的氨基酸疏水性;b. 肽段长度或相对分子质量;c. 序列信息;d. 二级结构信息[27],如 α-螺旋,β-折叠等。除此之外,预测肽段保留时间还要考虑:

(1)如何选取数据集。蛋白质组学研究目前常用多维 LC 分离技术[42],本文针对的是一维 RPLC 或多维 LC 中的反相分离部分,分离能力有限,适用于不是特别复杂的生物样品,所以能够鉴定到的肽段也不多。另外,蛋白质质谱鉴定实验的谱图鉴定率通常小于 20%,而鉴定结果中能够用于算法训练和测试的高可信肽段更少。前面提到的 Petritis 等[27]使用约 345 000 个肽段进行预测是特例,很难在其他实验室推广。根据实验数据的规模,按一定比例挑选高可信肽段的做法是比较可取的,如 Klammer 等[29]采用的方法。在确定比例时,要根据算法特点,防止数据集过小或拟合过度情况的出现。

(2)如何保证色谱行为的重现性。除了肽段自身的理化性质外,填料、色谱柱长度、流动相的组成、流速和外界温度等因素也对肽段的保留时间有重要的影响。此外,复杂的蛋白质组学样品会因不可逆吸附而改变毛细管色谱柱的分离性能,影响肽段保留时间在多次实验中的重现性。这就要求保留时间预测算法只能在一次独立实验中在线训练,并针对此次实验应用。基于这种思想,Klammer 等[29]应用基于 SVR 的预测算法,比较了多种实验条件下的预测效果,发现在各种条件下预测效果都不错。另外,Krokhin 等[43]最近以 6 个标准的肽段为基准,准确衡量和校正不同的 RPLC 条件对肽段保留时间的影响,该研究使得肽段在不同色谱条件下的保留时间具有可比性,这对促进肽段保留时间预测算法在蛋白质组学研究中的应用具有非常重要的意义。

(3)如何确定肽段保留时间。一般情况下,单一肽段的离子流色谱峰整体上相对平滑,整个肽段离子在这段时间内都有可能被鉴定到,可在肽段色谱峰上鉴定到的肽段母离子的对应时间为肽段保留时间,也可以取该肽段色谱峰最高点对应的时间作为肽段的保留时间,或者将所有鉴定出该肽段的谱图所对应的保留时间取均值。但现实实验中,有的肽段的离子流色谱峰分布非常不规则,所以需要专门建立模型来确定肽段的实际保留时间。张纪阳等[44]利用 3 次样条平滑的方法重构色谱峰,并且根据局部最小值的和来确定峰宽,以该峰宽内的局部最大点作为肽段的色谱保留时间。在通过该方法得到比较准确的肽段保留时间后,使用迭代最小二乘法进行肽段保留时间预测,最后预测保留时间与实际保留时间的相关系数可以达到 0.9031。他们还认为,对于复杂样品,保留时间因为本身波动较大,且可能存在肽段共洗脱问题,所以不能单独用来确认鉴定结果,但可以作为一种有效的辅助判据,排除假阳性鉴定结果

3　结论与展望

在基于质谱的蛋白质鉴定实验中,预测肽段在 RPLC 中的保留时间是一项非常有意义的工作。与

ML 法比,基于保留系数的 RC 法易实现,应用比较广泛,效果也不错,但基于 ML 的算法可扩展性强,具有广阔的发展前景。随着 RPLC 技术的发展,新的肽段保留时间预测算法不断产生,已有算法也在继续优化和改进,对保留时间的预测效果 R^2 已达到 0.93 左右[6]。但由于缺乏对这些繁复的算法进行比较和评价的统一标准,已有算法无法适应不同的实验要求和色谱条件,因此已成为制约保留时间预测广泛应用的瓶颈。除此之外,现在的预测算法大多针对常规的肽段,对修饰肽段的研究较少。如何在 PTM 研究中准确预测和应用肽段保留时间将是今后面临的重要课题。

虽然现在保留时间预测研究已取得一定的成果,但是还存在一些问题有待解决,离大规模实用还有一段距离。我们需要提高保留时间预测算法的准确性,建立各算法间统一的评价和比较标准,并比对不同色谱条件对保留时间的影响,开发更具普适性的算法,使保留时间预测真正成为蛋白质组学研究的重要手段之一。

参考文献:

[1] Aebersold R, Mann M. Nature, 2003, 422(6928): 198
[2] Domon B, Aebersold R. Mol Cell Proteomics, 2006, 5(10): 1921
[3] Beer I, Barnea E, Ziv T, et al. Proteomics, 2004, 4(4): 950
[4] Mchugh L, Arthur J W. PLoS Comput Biol, 2008, 4(2): 1371
[5] Zhang J F, He S M, Ling C X, et al. Rapid Commun Mass Spectrom, 2008, 22(8): 1203
[6] Spicer V, Yamchuk A, Cortens J, et al. Anal Chem, 2007, 79(22): 8762
[7] Sun R X, Fu Y, Li D Q, et al. Science in China Series E: Information Sciences (孙瑞祥, 付岩, 李德泉, 等. 中国科学 E 辑), 2006, 36(2): 222
[8] Draper N R, Smith H. Applied Regression Analysis. New York: John Wiley & Sons, Inc., 1998: 33
[9] Meek J L. Acad Sci USA, 1980, 77(3): 1632
[10] Browne C A, Bennett H P, Solomon S. Anal Biochem, 1982, 124(1): 201
[11] Guo D, Mant C T, Taneja A K, et al. J Chromatogr, 1986, 359: 499
[12] Guo D, Mant C T, Taneja A K, et al. J Chromatogr, 1986, 359: 518
[13] Casal V, Martin-Alvarez P L, Herraiz T. Anal Chim Acta, 1996, 326: 77
[14] Palmblad M, Ramstrom M, Markides K E, et al. Anal Chem, 2002, 74(22): 5826
[15] Palmblad M, Ramstrom M, Bailey C G, et al. J Chromatogr B, 2004, 803(1): 131
[16] Mant C T, Zhou N E, Hodges R S. J Chromatogr, 1989, 476: 363
[17] Mant C T, Hodges R S. J Chromatogr A, 2006, 1125(2): 211
[18] Tripet B, Cepeniene D, Kovacs D, et al. J Chromatogr A, 2007, 1141: 212
[19] Houghten R A, DeGraw S T. J Chromatogr, 1987, 386: 223
[20] Zhou N E, Mant C T, Hodges R S. Pept Res, 1990, 3(1): 8
[21] Krokhin O V, Craig R, Spicer V, et al. Mol Cell Proteomics, 2004, 3(9): 908
[22] Krokhin O V. Anal Chem, 2006, 78(22): 7785
[23] Krokhin O V, Cortens J P, Ghosh D, et al. Anal Chem, 2006, 78(17): 6265
[24] Dwivedi R C, Spicer V, Harder M, et al. Anal Chem, 2008, 80(18): 7036
[25] Shinoda K, Suqimoto M, Tomita M, et al. Proteomics, 2008, 8(4): 787
[26] Petritis K, Kangas L J, Ferguson P L, et al. Anal Chem, 2003, 75(5): 1039
[27] Petritis K, Kangas L J, Yan B, et al. Anal Chem, 2006, 78(14): 5026
[28] Shinoda K, Sugimoto M, Yachie N, et al. J Proteome Res, 2006, 5(12): 3312
[29] Klammer A A, Yi X, Maccoss M J, et al. Anal Chem, 2007, 79(16): 6111
[30] Oh C, Zak S H, Mirzaei H, et al. Bioinformatics, 2007, 23(1): 114
[31] Pfeifer N, Leinenbach A, Huber C G, et al. BMC Bioinformatics, 2007, 8: 468
[32] Eng J K, McCormack A L, Yates J R. J Am Soc Mass Spectrum, 1994, 5: 976
[33] Perkins D N, Pappin D J, Creasy D M, et al. Electrophoresis, 1999, 20(18): 3551
[34] Wang L H, Li D Q, Fu Y, et al. Rapid Commun Mass Spectrom, 2007, 21(18): 2985
[35] Strittmatter E F, Kangas L J, Petritis K, et al. J Proteome Res, 2004, 3(4): 760
[36] Sun W, Zhang L, Yang R F, et al. Rapid Commun Mass Spectrom, 2009, 23(1): 109
[37] Hsu Y R, Narhi L O, Spahr C, et al. Protein Science, 1996, 5(6): 1165
[38] Dasari S, Wilmarth P A, Rustvold D L, et al. J Proteome Res, 2007, 6(9): 3819
[39] Kim J, Petritis K, Shen Y, et al. J Chromatogr A, 2007, 1172(1): 9
[40] Baczek T, Sieradzka M. J Liq Chromatogr related technol, 2008, 31(16): 2417
[41] Baczek T, Wiczling P, Marszall M, et al. J Proteome Res, 2004, 4(2): 555
[42] Zhu G J, Liang Z, Zhang L H, et al. Chinese Journal of Chromatography (朱贵杰, 梁振, 张丽华, 等. 色谱), 2009, 27(5): 518
[43] Krokhin O V, Spicer V. Anal Chem, 2009, 81(22): 9522
[44] Zhang J Y, Zhu Y P, Xie H W, et al. Bulletin of the Academy of Military Medical Sciences (张纪阳, 朱云平, 谢红卫, 等. 军事医学科学院院刊), 2007, 31(1): 6