PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2015, 42(2): 101~114
www.pibb.ac.cn

“自顶向下 (top-down)”的蛋白质组学*
——蛋白质变体的规模化鉴定

孙瑞祥** 罗 兰 迟 浩 刘 超 贺思敏
(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；中国科学院大学，北京 100049)

摘要 高分辨率质谱技术的快速发展使得“自顶向下”的蛋白质组学(top-down proteomics)研究逐渐成熟起来. 在完整蛋白质水平上研究蛋白质组可以提供更精准、更丰富的生物学信息，特别是对于蛋白质上发生了多种关联性的翻译后修饰的情况. 另外，由于基因突变、RNA 可变剪接和大量蛋白质翻译后修饰的存在，同一个基因往往最终会产生多个 “蛋白质变体”(proteoform)，而要准确地鉴定这些蛋白质变体，也离不开“自顶向下”的蛋白质组学. 在蛋白质水平上的分离技术、质谱技术与生物信息学技术是完整蛋白质鉴定最关键的三项技术. 高效的分离技术是实现规模化蛋白质变体鉴定的前提，有效的质谱碎裂是提供可靠鉴定的核心，而快速准确的质谱鉴定算法则是数据分析效率的保障. 本文对这三项技术进行了详细总结，重点集中在生物信息学相关技术上，包括对完整蛋白质的质谱数据预处理、数据库搜索鉴定以及翻译后修饰定位等几个计算问题的讨论.

关键词 “自顶向下(top-down)”的蛋白质组学，串联质谱技术，生物信息学，蛋白质鉴定
学科分类号 Q51，TP39 **DOI**: 10.16476/j.pibb.2014.0078

蛋白质组学是 21 世纪初生物学的研究热点，其中，规模化的蛋白质鉴定是计算蛋白质组学中的一个基础问题[1]. 目前广泛采用的方法是称为“自底向上”(bottom-up，BU)的策略：通常先将蛋白质的复杂样品进行酶切产生肽段的混合物，然后通过液相色谱等技术将这些肽段的混合物进行分离，进而通过质谱技术将肽段碎裂，并根据碎裂谱图中的离子峰信息进行数据库搜索来鉴定肽段，最后将鉴定到的肽段进行组装、推理获得样品中所含有的蛋白质[2-3]. 所谓“底(bottom)”一般是指肽段，而“上(up)”则是指由肽段到蛋白质的推理过程，“自底向上”蛋白质鉴定策略的基本思想是“从局部推断整体”.

当前，蛋白质的 BU 鉴定技术正向着深度和快速覆盖全蛋白质组的方向发展，这主要源于质谱采集速度的大幅提升. 2013 年 10 月 Coon 研究组发表的论文[4]实现了一小时鉴定酵母的 3 977 个蛋白质，覆盖了酵母正常表达的约 4 500 个蛋白质的 88%，他们在文中定义了“深度覆盖”的含义为检测到正常表达蛋白质的 90%，这样，在蛋白质的鉴定数量上接近“深度覆盖”酵母的全蛋白质组；然而，仍有 14%的蛋白质(538 个)仅鉴定到了一个肽段，鉴定蛋白质的序列覆盖率平均也仅有 18%，这说明采用“自底向上”以肽段为研究中心的方法鉴定到的蛋白质序列覆盖率很低，大约存在 10%～20%鉴定到的蛋白质，“仅窥独树，未览森林”. 在这种情况下，当需要从鉴定到的肽段完成蛋白质的组装时，由于蛋白质变体(proteoform[5])之间在氨基酸序列上的高度相似性，往往会导致蛋白质推理结果的不确定性[6]，其本质是由于 BU 方法天然的信息缺失所导致的. 另外，翻译后修饰之间的关联信息往往对生物过程的调控起着关键作用，如果一个蛋白质上同时发生了多个翻译后修饰，BU 分析方法中的酶切过程使得肽段与蛋白质之间的归属信息缺失，导致要分析这些不同修饰之间的

* 国家重点基础研究发展计划(973)(2013CB911203, 2010CB912701)资助项目.
** 通讯联系人.
Tel: 010-62600822, E-mail: rxsun@ict.ac.cn
收稿日期：2014-03-19，接受日期：2014-06-27

关联变得困难．上述 BU 技术上的局限性阻碍了“自底向上”蛋白质组技术的深入应用．

为了缓解 BU 技术上的局限性，“自中向下”(middle-down，MD) 的蛋白质组技术采用了不同的生物酶，它可以获得较长的肽段，如 6.3 ku 以上的肽段[7]．这样，MD 可以分析鉴定较长的肽链上同时发生的几个翻译后修饰，相比 BU 方法来说，它可以分析的肽段范围更广，在多泛素化链结构[8]和单克隆抗体[9]等分析上获得了应用．

“自顶向下”(top-down，TD) 的蛋白质组技术则不再需要酶切的过程，直接以完整的蛋白质为分析对象，它可以提供完整蛋白质更精准、更丰富的生物学信息，从根本上解决了上述 BU 中存在的问题[10−16]．所谓“顶(top)”是指对完整蛋白质分子质量的准确测定，而“下(down)”则是指通过串联质谱技术实现对完整蛋白质的碎裂．这样，通过完整蛋白质的质量及其碎裂谱的信息可以实现真正意义上的蛋白质鉴定，它的序列覆盖率可以达到 100%[10]，而基于 BU 策略的主流技术覆盖率则小于 20%[11]，也就是说，蛋白质还有超过 80%的序列部分处于未鉴定状态．相比而言，TD 能够保留多种翻译后修饰之间的关联信息，因而在蛋白质组学的研究中逐渐成为与 BU 和 MD 技术优势互补的、前景看好的新技术方向．

基于学科发展的需要，本文对“自顶向下”的蛋白质组学分别从分离技术、质谱技术与生物信息学技术等三个方面进行了详细的阐述，并对当前生物信息学中质谱分析的计算问题进行了专门分析，最后给出了其面临的技术挑战和未来的发展趋势．

1 “自顶向下”的蛋白质组学

人类蛋白质组项目(human proteome project)的目标是确定组成人体的所有蛋白质分子，以更好地理解和检测各种疾病，以及为药物治疗而制定一个蛋白质的参考列表[11]．随着质谱技术的快速发展，人类蛋白质的鉴定数目不断被刷新，在 20 128 个蛋白质编码基因中，从质谱数据、抗体、氨基酸测序和 3D 结构等方面得到验证的蛋白质达到了 15 646 个，除去 638 个可疑的(dubious)，仍有 3 844 个蛋白质目前还没有实验上的证据[17]．要寻找到这 3 844 个“丢失的蛋白质(missing proteins)”，高度同源的蛋白质需要考虑，因为当前基于质谱技术的鉴定策略多是采用基于酶切的方式，鉴定到的蛋白质是以“组(group)”的方式呈现的，这样，当鉴定到的肽段同时匹配上几个高度同源的蛋白质时就无法确定哪些蛋白质是真实存在的，这样的现象在人类蛋白质中大量存在[17]．以完整蛋白质为研究对象的“自顶向下”的蛋白质组学，为解决这个问题提供了新的思路和技术实现的可能，这主要也是源自最近几年内质谱技术的快速发展．

早期的质谱技术主要用途是测量元素的同位素，之后出现了可以测量小分子的质谱．20 世纪 80 年代末电喷雾离子化(electro-spray ionization，ESI)与基质辅助激光解吸附离子化(matrix-assisted laser desorption and ionization，MALDI)两项软电离技术的发明，使得质谱技术可以在测量生物大分子上得到广泛的应用，以肽段为中心的质谱分析已经成为当前蛋白质鉴定的常规手段．随着高分辨率质谱技术越来越成熟，直接测量完整蛋白质的分子质量与蛋白质的碎裂谱成为可能．质谱技术的发展还可以直接测量蛋白质复合体的质谱．从上述质谱技术的发展过程来看，可测量的对象由小逐渐变大，从元素到小分子，从肽段到蛋白质，再到蛋白质的复合体，这反映出了质谱技术发展的趋势，同时也反映出了蛋白质研究的技术需求．

最近几年，随着用于完整蛋白质分析的质谱技术越来越成熟，特别是以轨道回旋离子阱(orbitrap[18−21])质量分析为核心的质谱仪器的进步，越来越多的研究集中到完整蛋白质的质谱分析上．图 1 给出了查询谷歌学术网(Google Scholar)在 2000～2013 年发表论文中主题词包含“top down mass spectrometry”的数量，特别是 2007 年后增长明显，这说明质谱技术在 TD 研究中起着非常关键的作用．

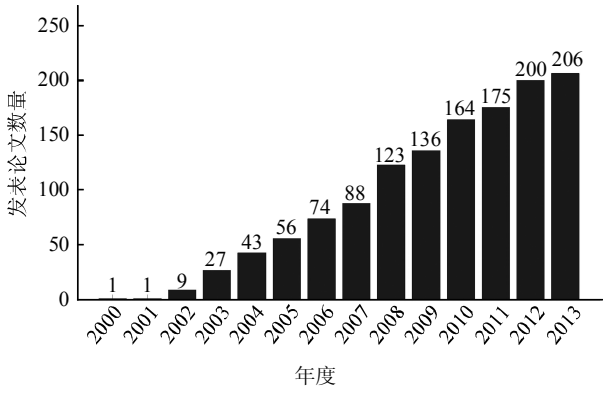

Fig. 1 Publications of top-down mass spectrometry (2000～2013)

图 1 Top-down 质谱技术的年度发表论文数量 (2000～2013)

最近两年，在国际学术界围绕“自顶向下”蛋白质组学的研究交流十分活跃．2012 年 3 月国际上成立了 TD 联盟(consortium for top down Proteomics)，该组织的使命是促进完整蛋白质综合分析的革新研究、合作与教育[22]．在 2013 年 3 月出版的 *Nature Methods* 期刊上，该组织联合发表了“蛋白质变体(proteoform)”的定义，它用于表征由同一个基因形成的所有不同的蛋白质分子，主要来源有基因突变、可变剪接和翻译后修饰等[5]．2013 年 1 月 24 日至 27 日在美国的 Florida 召开了第 25 届美国质谱学会 Sanibel Top-down 质谱技术国际会议，吸引了来自 16 个国家的 135 位研究人员参加，会议包括 26 个邀请报告、8 个墙报介绍的短报告和 34 个墙报展出，会议内容覆盖了完整蛋白质分析的三个主要方面，即分离技术、质谱技术与生物信息学技术[15, 23]．另外，在 2013 年 6 月，国际期刊 *Analytical Chemistry* 和 *Journal of Proteome Research* 联合出版了关于 top-down 蛋白质组学的虚拟专辑(Virtual Issue)，汇总列出了之前两个期刊曾出版过的 64 篇论文，包括 1 篇综述、29 篇分离技术、28 篇质谱技术和 6 篇蛋白质鉴定信息学方面的论文[24]．2013 年的美国质谱年会上也有 19 个关于 TD 技术方面的专场报告[25]．所有这些都显示出 TD 技术最近在快速发展，在解决具体的生物学问题中也开始发挥出不可替代的作用[26-32]．一个成功的应用例子，如文献[33]采用了 TD 技术和质谱成像技术相结合，研究了神经发育过程出现失调的蛋白质标记物，发现有 22 个蛋白质发生了相对丰度的变化，进而对它们的功能进行了进一步的分析．目前，国际上也已有关于 TD 技术相关的综述论文发表，如文献[34-37]，文献[38-40]则专门讨论了关于大蛋白质的鉴定问题．这些文献在分离技术和质谱分析两个方面论述较多，本文归纳了这些文献的内容，并在生物信息学方面进行了较为详细的讨论．另外，国内还未见系统综述最新 TD 技术的相关文献发表．

一般地，TD 技术主要包括完整蛋白质的分离技术、质谱技术和生物信息学技术三个方面，如图 2 所示．前端的蛋白质分离主要是为了降低质谱分析进样的复杂度，提高质谱数据的“纯度”和信噪比，高效的分离技术是实现规模化蛋白质变体鉴定的前提．质谱分析主要包括测量完整蛋白质的准确分子质量和获得蛋白质的碎裂谱图两个方面，其中最关键的质谱碎裂是可靠鉴定打分算法依赖的核心．后端的生物信息学分析必须依靠高效的算法和软件来完成，蛋白质变体的鉴定目前主要依靠数据库检索的方法来实现，所以快速准确的蛋白质质谱检索算法就成为获得最终蛋白质变体鉴定结果的保障．最近 2～3 年，这些技术发展迅速，尤其是分离技术和质谱技术的发展使得可以分析的蛋白质规模得到了大幅度提升．目前 TD 技术已经可以实现上千个蛋白质的鉴定[41-42]，并逐渐变为现实可用的技术[14]．下面将分别具体介绍这些技术的发展．

Fig. 2　Three main techniques for top-down proteomics

图 2　Top-down 蛋白质组学中的三项主要技术

2　完整蛋白质的分离技术

早期的 TD 研究主要集中在纯化的单个蛋白质或者蛋白质复合体上，由于这些样品的复杂度不高，在蛋白质层次上的分离方法研究与应用并不多．随着分析样品的复杂度增加，特别是对全蛋白质组的复杂样品分析，在质谱分析之前进行有效的蛋白质分离可以显著地降低样品的复杂度，提高质谱分析的效能，实现高通量的蛋白质鉴定．因此，当前对蛋白质分离技术的研究已经成为 TD 中最活跃和最基础的部分，也是技术进展最显著的部分，而且它已经使得分析整个蛋白质组成为现实．其实，已经广泛使用的二维凝胶电泳本身就是对蛋白质混合物的分离，但由于其回收率低、实验操作耗时耗力、重复性差等因素，在 TD 研究中并没有取得很好的使用效果．在 BU 中广泛采用的液相色谱分离技术也主要是适用于肽段的有效分离，在完整蛋白质水平上的直接分离应用效果也并不理想，其中的主要原因在于蛋白质与肽段在理化特性上的差异．因此，寻找适用于蛋白质、并且能够实现高通量地有效分离就成为 TD 研究中的一个基础问题．

目前，最有效的蛋白质分离方法主要分为离线(off-line)与在线(on-line)两种方式，前者以 Kelleher

实验室发展起来的“四维离线分离”方法[41-42]为代表；后者以 Ljiljana 实验室发展起来的弱阳离子交换(weak cation exchange，WCX)结合亲水交互液相色谱(hydrophilic interaction liquid chromatography，HILIC)在线分离方法[43-44]为代表.

2.1　离线分离技术

Kelleher 实验室发展起来的“四维离线分离”平台如图 3 所示[41]. 首先蛋白质混合物进行等电位点聚焦(isoelectric focusing，IEF)，它主要依据蛋白质的等电位点进行分离(图 3a)，与其他分离方法相比，它具有更高的分辨率和分离精度. 接下来上述分离物进行凝胶洗脱液相分离截留电泳(gel elution liquid-based fractionation entrapment electrophoresis，GELFrEE)，依据蛋白质的分子质量进行分离(图 3b)，具有更高的样品回收率，可以很好地与其他分离方法相连接. 这两种分离方法的结合，再加上 LC-MS(图 3c)构成了完整的四维蛋白质分离平台，显著提高了蛋白质分离的通量. 在对人类 HeLa 细胞蛋白质组的分析中鉴定到了 1 043 个蛋白质(对应 3 093 个蛋白质变体)，在鉴定数量上较之前的最好结果提升了 10 倍. 最近在 H1299 细胞蛋白质组的分析中鉴定到了 1 220 个蛋白质(对应超过 5 000 个蛋白质变体)[42]. 四维离线分离平台真正实现了组学规模上的完整蛋白质鉴定，解决了之前困扰 TD 的低通量问题，代表了 TD 研究中分离技术上的一个里程碑.

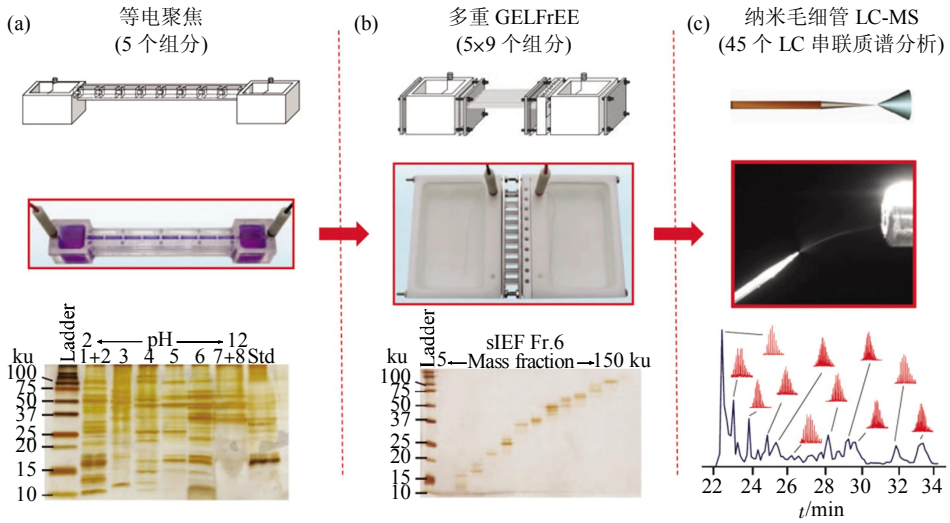

Fig. 3　Four-dimensional separation platform for intact proteins[41]

图 3　完整蛋白质的四维分离平台示意图[41]

(a) 等电位点聚焦分离. (b) 多重 GELFrEE 分离. (c) RPLC 与质谱分析.

2.2　在线分离技术

虽然离线分离技术可以实现高通量的蛋白质鉴定，但是由于分离出来的组分(fraction)需要手动收集，效率和自动化程度较低，因此，发展在线的高效分离方法是提高分离效率的另一条技术路线. 在线分离方法的代表是以 RPLC 和 WCX-HILIC 相结合的分离技术，如图 4 所示[43]. WCX 主要依据蛋白质的静电荷多少进行分离，它由蛋白质分子的 p*I* 和溶液的 pH 共同决定，在阳离子交换柱上，只有 p*I* 大于流动相 pH 的蛋白质在柱子上保留，而 p*I* 小于或等于流动相 pH 的蛋白质不保留，即使它们的 p*I* 值彼此之间有差异，也都作为溶剂峰同时被直接冲洗出来. HILIC 可以被看作是正相色谱向水性流动相领域的延续，其流动相是水相缓冲液及有机溶剂，固定相是强亲水性的极性吸附剂，如硅胶键合相、极性聚合物填料或离子交换吸附剂. 这些固定相的共同特点是它们和水的作用力很强，因此属于“亲水性”. 由于使用水溶性有机溶剂、高有机相流动相、低的缓冲盐浓度、不使用离子对试剂、与质谱检测良好的兼容性和灵敏度等优点，应用方面也越来越广，特别是在发生翻译后修饰的蛋白质分离上逐渐被广泛采用.

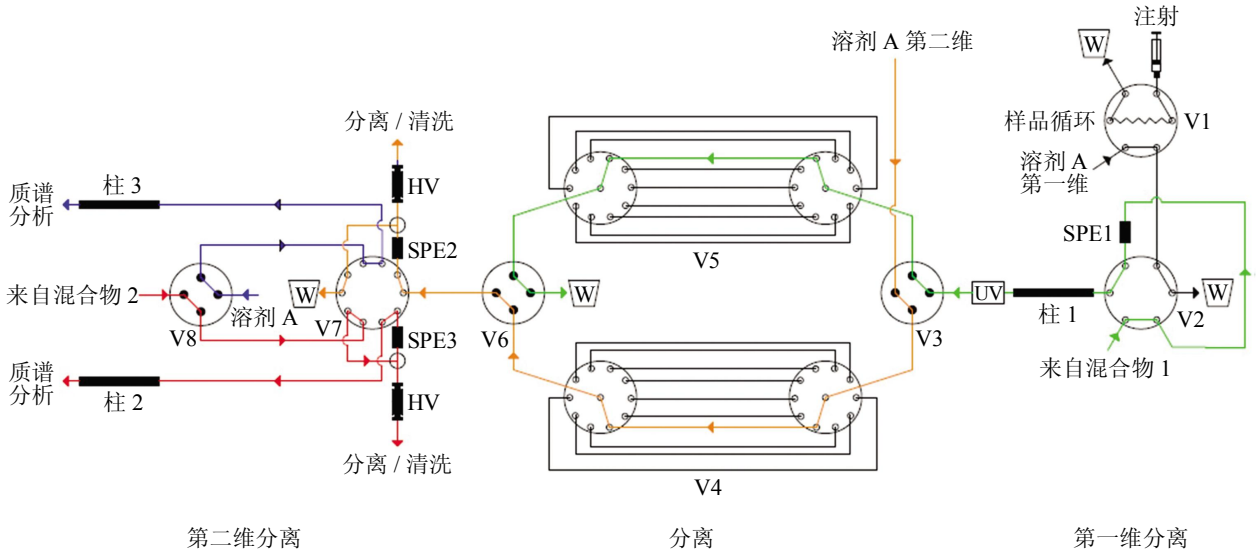

Fig. 4　On-line separation platform based on RPLC/WCX and HILIC[43]

图 4　RPLC/WCX 和 HILIC 在线分离平台示意图[43]

V：Valve(阀)；W：Waste(废弃物)；MS：Mass spectrometer(质谱仪).

在图 4 分离平台右侧的第一维分离中，V1和 V2 两个纳米流量阀用于样品注入 SPE1 柱和分离柱 1(RPLC). V3、V4、V5 和 V6 四个阀用于第二维分离. 第一维的分离组分被载入 SPE2 和 SPE3 两个柱，然后分别用两个 WCX-HILIC 柱来进行第二维的分离[42]. 此分离平台适用于复杂的翻译后修饰蛋白质的分离，如组蛋白(histone)，文献[44]中采用该方法鉴定到了 708 个组蛋白的变体，实现了高灵敏度和高通量鉴定复杂的翻译后修饰蛋白质.

在 TD 分离技术中，除了上述介绍的两类方法外，还有毛细管电泳(capillary electrophoresis，CE)和强阳离子交换(strong cation exchange，SCX)等，由于应用较少，这里就不再赘述，文献[37]中总结了目前常用的各种蛋白质分离技术.

一个高效的蛋白质分离系统，特别是分析复杂的蛋白质样品，如人类蛋白质组，可以大大降低质谱分析的负担，减小进入质谱仪的蛋白质“拥堵程度”，人们追求的理想分离效果是使得要分析的蛋白质能够“排队”依次进入质谱仪中进行分析. 然而，由于分离技术的局限性和样品的复杂度所限，现在还没有这样理想的分离平台，还需要进一步研究开发更好的分离技术，因为高效的蛋白质分离技术是实现规模化蛋白质变体鉴定的前提.

3　完整蛋白质的质谱技术

与肽段的质谱分析相比，蛋白质的质谱分析难度更大，也更依赖于质谱技术的进步，对于每一个完整的蛋白质分子，利用质谱技术获得其准确的分子质量和信息丰富的串联质谱是可靠鉴定的基础，因此，TD 质谱技术的研究主要包括两个方面：a. 利用高分辨率质谱技术测量完整蛋白质的准确分子质量；b. 利用高精度的串联质谱技术获得完整蛋白质的碎裂质谱图.

3.1　蛋白质分子质量的准确测定

完整蛋白质的准确分子质量对于蛋白质的鉴定和翻译后修饰分析可以提供判定信息，因此，在 TD 质谱分析中首要的任务是获得蛋白质的准确分子质量. 其实，两种软电离技术 ESI 和 MALDI 的发明就是针对生物大分子，如蛋白质分子. ESI 可以获得带多个电荷的蛋白质离子，它在一级质谱分析中可以清晰地显示出同一个蛋白质分子因带不同的电荷状态而形成不同的同位素峰簇. 图 5 给出了一个实例，同一个蛋白质分子，所带电荷数可清晰地看到从+7～+19，共 13 种不同的电荷状态，每个谱峰(如果局部放大，实际为系列同位素峰)上都标注出了电荷数. 如果把某个电荷状态对应的谱峰局部放大，如+15，则可以看到图中箭头右侧的放大图，在高分辨率质谱仪上可以清晰显示出系列同位素峰. 由于同一个蛋白质变体具有同一个质量，这样根据图中不同电荷状态对应的谱峰信息，就可以推断出该蛋白质变体的分子质量，这个过程一般称为“去卷积(deconvolution)[45]”.

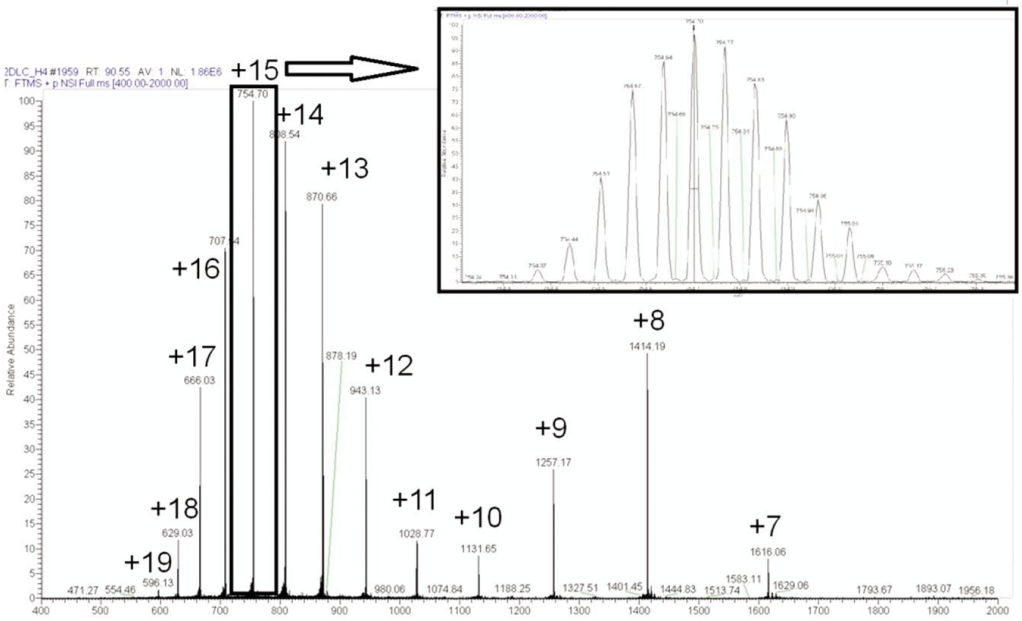

Fig. 5　An example of MS1 from an intact protein

图 5　完整蛋白质分子的一级质谱图实例

右上角的局部放大图对应的是带 15 个电荷该蛋白质分子的同位素峰簇，数据来自文献[44].

3.2　蛋白质的串联质谱技术

串联质谱数据中包含了用于推断序列的氨基酸信息，因此，如同鉴定肽段的串联质谱作用，完整蛋白质的串联质谱数据也是序列鉴定的基础．不过，由于蛋白质的分子质量一般要比肽段大得多，如何使蛋白质分子在质谱仪中碎裂而产生信息丰富的、利于鉴定的串联质谱数据在技术上则具有较大的难度，存在的主要问题包括碎裂能量的选择、多电荷谱峰的处理、检测灵敏度的提高等等．早期的蛋白质碎裂质谱技术主要使用低能量的碰撞诱导裂解(collision induced dissociation，CID)．虽然对于大多数肽段而言，CID 可以获得丰富的碎裂信息，但对于完整的蛋白质分子，CID 碎裂得并不充分，主要是在有限的少数几个肽键处断裂，如氨基酸 P 的 N 端，或氨基酸 D 和 E 的 C 端等．这样，如果蛋白质上发生了翻译后修饰，低能量的 CID 谱图往往不足以确定这些修饰发生的位点，如发生磷酸化或糖基化的蛋白质．McLafferty 等发展起来的电子捕获裂解(electron capture dissociation，ECD)技术更适合用于完整蛋白质的碎裂，它克服了 CID 的不足．ECD 是采用自由的热电子与多电荷的蛋白质离子相互作用，作用过程中产生蛋白质分子的断裂，主要产生以 c 和 z 为主导的碎片离子，由于 ECD 是非各态遍历(non-ergodic)的，它可以保留发生修饰的基团在碎片离子上，相比 CID，ECD 可以准确判断修饰的位点[46-47]．因此，在 TD 研究中获得了一些重要的应用[47-48]．不过，由于 ECD 主要是配置在价格昂贵的傅里叶变换离子回旋共振(fourier transform ion cyclotron resonance，FTICR)质谱仪上，其推广应用受到了仪器方面的局限．Hunt 等针对这个问题，研制了与 ECD 机理类似的电子转运裂解(electron transfer dissociation，ETD)技术，与 ECD 所不同的是 ETD 是通过携带电子的阴离子与肽段或蛋白质离子相互作用，它可以在应用更普及的离子阱型质谱仪上实施[49]．ETD 不仅在肽段鉴定，特别是发生翻译后修饰肽段鉴定上得到了应用[50-51]，而且在完整蛋白质的鉴定上也得到了很好的应用效果[52-55]．ECD 和 ETD 成为完整蛋白质碎裂的首选方法．另外，最近几年发展起来的高能碰撞裂解(high-energy collision dissociation,HCD)也开始应用于完整蛋白质的碎裂[56].

3.3　用于完整蛋白质分析的质谱仪

在质谱仪器方面，有多种不同类型的质谱仪曾应用在 TD 研究中，包含了常用类型的质谱仪器，

文献[10]中给出了早期使用的质谱仪．虽然高分辨率、高精度的质谱仪在蛋白质分子质量的测量方面具有精度方面的优势，但根据上述图 5 中所示的蛋白质完整分子 ESI 一级质谱的特点，对于低分辨率的质谱仪，也可以推算出蛋白质的分子质量(虽然同位素峰不足以确定电荷)．当然，现在的 TD 研究还是更多地使用高分辨率的质谱仪，主要原因是它能够解析蛋白质的同位素谱峰，进而确定电荷数和分子质量．特别是最近几年，越来越多的 TD 研究开始使用 Orbitrap 系列的质谱仪[18−21, 42, 44]．虽然 Orbitrap 质谱仪在 2005 年才开始商业化，但其在蛋白质组学中的应用已经普遍．Orbitrap 具有 FTICR 级别的测量精度而又不需要超导磁场，这使得它的使用和维护更加方便．同时，它还可以与其他更灵敏的质量分析器结合起来组成混合的质谱仪，如 LTQ-Orbitrap 等．TD 的研究更依赖于高精度的质谱仪，随着新一代 Orbitrap 仪器的开发，其分辨率和测量速度得到了进一步地提升，已经成为 TD 研究中的主力仪器[20, 42, 57]．

4　完整蛋白质分析的生物信息学技术

当前，TD 研究中面临的最大挑战不是分离技术，也不是质谱技术，而是质谱数据分析的算法与软件，即生物信息学技术[58]．前文介绍的 TD 研究出版虚拟专辑[24]中，分离技术和质谱技术的论文分别占 46%和 44%，而生物信息学方面的论文还不到 10%，这也说明亟需开展生物信息学方面的研究．与肽段的质谱数据处理相比，蛋白质的质谱数据处理技术的挑战主要体现在蛋白质的谱图更加复杂，数据库检索的规模更大，缺少有效的算法和数据处理软件．

与肽段的谱图相比，完整蛋白质的谱图更加复杂，质谱数据处理起来也更加棘手，这源于在蛋白质的一级和二级谱图中，高电荷离子比较典型(如图 5 的一级谱)，而在常用酶切的肽段谱图中，大多数离子不超过 3 个电荷．另外，由于蛋白质分子及其碎裂后的离子质量范围更宽，大质量离子的同位素峰数更多．由于质谱仪的测量范围一般是固定的，大量高电荷、高质量离子的谱峰出现导致谱峰的混叠现象更普遍，使得谱图的解析更加困难．目前，在肽段的谱图处理上已经有很多的软件可选用，而在蛋白质质谱数据解析上则缺少有效的软件，很多时候不得不依赖手工处理和校验．

除了蛋白质的质谱数据本身表现出的复杂性，在数据库检索方面也面临着巨大计算量的挑战．在蛋白质鉴定中，我们需要考虑蛋白质不同的变体形式，如何根据蛋白质的序列生成要匹配的候选变体成为不同于肽段搜索的一个计算难题．一般地，在肽段鉴定中，我们容许肽段上发生的修饰数目较少，如 3 个修饰位点，这主要是由于肽段序列的平均长度相对蛋白质要短很多．然而，对于完整的蛋白质，发生翻译后修饰等变化的位点数目可能要多很多，这样，我们容许一个蛋白质序列上的可变修饰位点数就会增加，例如 10～20 个，再考虑到每个位点上可能修饰类型的不同，组合起来容易产生巨量的蛋白质变体形式．以组蛋白 H4 为例，虽然这个蛋白质只有 102 个氨基酸，是一个小蛋白，但在这个蛋白质的靠近 N 端区域，经常会发生多种翻译后修饰的组合，特别是发生在 K 和 R 上的就可能有乙酰化、甲基化、二甲基化、三甲基化等等．通过 UniProt 数据库提供的信息预测出的 H4 变体形式理论上就达到了 260 亿的规模[59]，如何产生、存储和管理这么大规模的变体形式，需要高效的生物信息学算法等技术的支持．

在生物信息学技术方面，类似于肽段的质谱鉴定方法，完整蛋白质的鉴定主要也是采用数据库搜索的方法．因此，如何衡量蛋白质质谱与蛋白质变体匹配的好坏成为设计和实施一个新的蛋白质搜索引擎的核心．在搜索数据库之前，一般需要将蛋白质的原始质谱数据转换为一个标准的谱峰列表(如单电荷、单同位素质量)，这样便于后续程序的处理，这个转换过程称为蛋白质质谱数据的预处理．在肽段的高分辨率谱图中，单同位素峰一般较容易识别；而在蛋白质的高分辨率谱图中，特别是大质荷比的离子，单同位素峰往往不显著，需要通过估算判断，在信噪比较低或信号干扰的情况下，容易产生估算偏差，这是完整蛋白质质谱数据预处理时需要专门考虑的问题．

4.1　完整蛋白质质谱数据的预处理

2000 年 Horn 等[45]开发了针对生物大分子的质谱数据预处理算法 THRASH，采用了一种“减法式谱峰发现(subtractive peak finding)”的方法来寻找可能的同位素峰簇，进而通过最小方差拟合理论同位素峰与实验同位素，确定最高谱峰的质量．另外，他们还提出了一种新的计算谱峰信噪比的方法，其基本思想是利用所有谱峰的强度统计直方图，认为噪音峰的强度分布应是最密集，对应于统计直方图的最高点，由此确定了噪音的基线后，重

新定义信噪比 *S*/*N* 为：

$$\frac{S}{N}=\frac{I_{\text{peak}}-I_{\text{baseline}}}{\text{FWHM}}$$

其中 I_{peak}，I_{baseline} 分别为谱峰的强度和噪音基线的强度，FWHM(full-width half maximum)为噪音峰的宽度，采用分布的半峰宽．THRASH 算法已经实现在很多的软件系统中，成为 TD 质谱数据预处理的主要算法．虽然 THRASH 算法已经被广泛应用，但随着质谱数据规模的增大，其计算速度比较慢也成为一个未解决的问题，这主要是由统计密度函数的计算本身需要比较大的计算开支所导致的．另外，THRASH 算法对于重叠的同位素谱峰处理效果也不够理想．

2010 年 Liu 等[60]开发了针对完整蛋白质的质谱数据预处理算法 MS-Deconv，采用了组合算法来解决去卷积的问题，特别是对于同位素峰簇重叠或部分谱峰叠加的情况．它首先产生大量的候选的同位素包络(isotopomer envelope)，然后用图(graph)来表示这些候选同位素包络间的关系，通过动态规划算法寻找图中的最优路径，获取最高打分的同位素包络，最后提取对应的单同位素峰质量，完整流程如图 6 所示．与其他处理方法不同，MS-Deconv 处理的不是单个包络，而是批量的包络，这样更容易获取最优的包络．在具体数据上的性能评测比较中，MS-Deconv 比 THRASH 多识别出了 20%的谱峰，速度提升了 33 倍[60]．

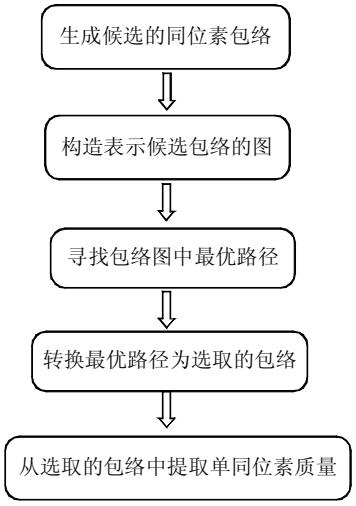

Fig. 6 Flowchart of MS-Deconv[60]

图 6 MS-Deconv 的流程图[60]

4.2 完整蛋白质的数据库搜索鉴定算法

在蛋白质数据库搜索鉴定算法中，主要有 ProSight 和 MS-Align+，是采用了不同的方式处理发生翻译后修饰的蛋白质鉴定问题．

ProSight 是第一个 TD 质谱鉴定算法，采用了如下的泊松分布概率打分模型[61]：

$$P_{f,n}=\frac{(xf)^n\times c^{-xf}}{n!} \tag{1}$$

$$x=\frac{1}{111.1}\times 2\times(M_a\times 2) \tag{2}$$

$$x'=\frac{1}{111.1}\times 2^{\psi+1}\times(M_a\times 2) \tag{3}$$

其中，$P_{f,n}$ 表示串联质谱中 n 个碎片离子中有 f 个离子随机匹配上了谱峰的概率，x 表示碎片的某一个质量在平均意义下匹配上某个碎裂离子的概率，计算方法如公式(2)所示，氨基酸的平均分子质量为 111.1 Da，匹配离子类型考虑两种(CID 或 HCD 的 b 和 y 离子，ECD 或 ETD 的 c 和 z 离子)，M_a 为质量匹配的容差，其对应的匹配窗口是长度为 2 倍的 M_a，这样就得到了计算公式(2)．如果还考虑 ψ 种修饰，那么每种修饰会增加 2 种新的离子谱峰，分别为 b+修饰，y+修饰(CID 或 HCD 碎裂谱)，这样，公式(2)就调整为公式(3)．有了上述的概率模型，对于实际的匹配离子数目就可以带入公式(1)中计算蛋白质与谱图随机匹配的概率 $P_{f,n}$ 分数，这是 ProSight 核心的打分部分．文献[50]中还对此打分模型的适用性在理想的情况下进行了分析，结果表明，在搜索 5 000 个蛋白质的数据库时，如果离子的匹配误差是±0.1Da，那么只要匹配上 3～4 个碎片离子，鉴定的可靠度就可以达到 99.8%．

ProSight 算法搜索数据库查询蛋白质的方式包括如下三种[62-64]：

a．通过完整蛋白质的分子质量；

b．通过谱图获得的序列片段；

c．通过上述两种方法的结合．

在生成蛋白质变体的虚拟数据库(virtual database)上，ProSight 采用了一种称为“鸟枪法标注(shotgun annotation)”的方法，其实质上是根据数据库中已有的标注信息自动枚举生成蛋白质变体的所有可能形式，比如突变、可变剪接和翻译后修饰等，形成与蛋白质质谱数据进行匹配的虚拟数据库．这种方法减轻了蛋白质变体与其质谱匹配打分的负担，但它的局限性是可扩展能力弱，如果变体形式来源太多，则容易导致“组合爆炸”的现象．如组蛋白 H3.1 理论上就可能有 40 万亿个不同的变体；一个 4M 大小的蛋白质数据库，变体形式需要的存储空间能达到 30G，这么大规模的变体使用标

注数据库的方式全部生成在实施中已经变得不可行，而且该方法在搜索时对于注释数据库中不存在的变体形式则无能为力．这正如美国加州大学圣地亚哥分校 Pevzner 教授等所指出的，这是 ProSight 的“阿喀琉斯之踵(achilles heel)”[65].

Pevzner 教授研究组开发了 MS-Align+[66]算法，是采用了与 ProSight 不同的思路．MS-Align+算法借鉴了之前他们在肽段鉴定中开发的谱图比对(spectral alignment，SA)方法，将其扩展到了完整蛋白质的鉴定上，着重要解决三个问题[67]：

a．之前在肽段鉴定上，SA 一般处理肽段上 1～2 个修饰，然而，在蛋白质上，可能需要处理 10～20 个修饰.

b．蛋白质鉴定上经常处理同一个蛋白的不同变体形式，这样，在一张谱图中可能含有多个变体形式需要鉴定，类似于肽段鉴定中的混合谱问题，但这比混合谱更难，因为不同变体在质谱中多数碎片离子的质量都相同.

c．考虑 TD 谱图的特点，如碎裂不完全的问题，质量测量误差问题等.

MS-Align+采用了动态规划的算法实现谱图与序列的比对，可以处理未知修饰的鉴定，不需要像 ProSight 那样预先采用“鸟枪法标注”产生所有的候选变体，而是根据实际谱图提供的信息优化获得最优的变体，这样，就在时间和空间上提高了鉴定的效率．与 MS-TopDown[67]相比，MS-Align+通过优化动态规划算法显著提高了计算速度．另外，对鉴定结果提供了统计显著性的评价指标(E-Value).

针对发生多翻译后修饰的蛋白质鉴定问题，MS-Align-E[59, 68]算法同时考虑了未知修饰和已知修饰的鉴定，与其他算法相比，在组蛋白 H4 上鉴定到了更多的变体形式．然而，动态规划算法本身的计算复杂度在同时考虑未知修饰和已知修饰时仍然较高，所以，容许的修饰类型和位点数目仍然受限.

在 TD 质谱数据的鉴定方面，除了 ProSight 和 MS-Align+算法，最近还有其他几个相关算法发表，如 PIITA[69]、ProteinGoggle[70]等，由于这些算法还仅局限在实验室中研究应用，这里不再具体介绍.

4.3　完整蛋白质翻译后修饰的定位

同酶切肽段上发生的翻译后修饰一样，在完整蛋白质上发生的翻译后修饰也需要定位．如果翻译后修饰能够被准确定位到某一个确定的氨基酸上，将对其生物功能的研究提供重要的信息，所以，对于翻译后修饰定位的生物信息学研究也是蛋白质鉴定中的一个重要子问题．虽然在肽段上的翻译后修饰定位有不少算法，如磷酸化的定位算法，然而，不同碎裂方法需要采用不同的分数来控制误定位率[71]．由于蛋白质的长度一般较肽段要长得多，发生的翻译后修饰类型和位点数目也随之增多，如何准确定位修饰主要取决于用于确定位点的碎裂离子峰是否显著，如图 7 中的三个蛋白质变体 PF1、PF2 和 PF3，能够区分甲基化是发生在第 3 个氨基酸 R 上(PF1、PF2)还是第 5 个氨基酸 K 上(PF3)，主要取决于 b3 和 b5 离子谱峰(或对应的 y 离子)信号，当这些辨别离子谱峰缺失或信号很弱的情况下，就无法通过质谱数据提供的信息来实现准确定位．而在蛋白质变体之间，如果发生的翻译后修饰类型和位点数目都相同，只是修饰位置不同，那么要鉴定到某个或某几个变体，必须能够准确定位到这些修饰才能唯一确定变体形式，如图 7 中的三个变体上都发生了 3 个修饰：甲基化、乙酰化和二甲基化，这样，它们的母离子质量(蛋白质变体的质量)是完全相同的，无法在一级质谱上区分，只能通过二级谱图来确定存在变体的形式，这样，能够确定修饰定位的离子就很关键，只有存在这些离子的情况下，我们才能确定是鉴定到哪个或哪些变体．目前的技术方法还没有很好地解决这个问题，软件提供的鉴定结果主要是蛋白质的信息，对于是否准确定位还没有足够好的分析方法.

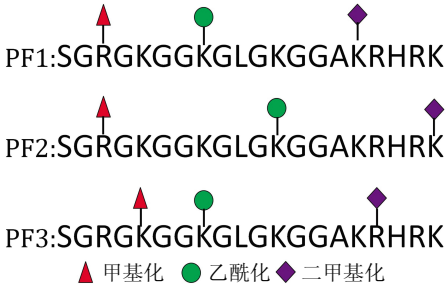

Fig. 7　Three proteoforms with different PTM combinations

图 7　不同的修饰定位组合辨别不同的三个蛋白质变体

4.4　完整蛋白质鉴定软件的性能评估

为了评估现有 TD 软件的鉴定性能，我们在测试数据上比较了三款软件：我们课题组开发的

pTop、前文介绍过的 ProSight 和 MS-Align+. 测试数据集来自美国 PNNL 的 EMSL 实验室产生的人类组蛋白 H4 的质谱数据[44]，采用的质谱仪为 LTQ-Orbitrap-Velos，一级谱和二级谱分辨率均为 60 000. 对同一个母离子分别采用 CID 和 ETD 的碎裂方式得到 2 张二级谱. H4 数据集一共有 2 698 张二级谱，其中 1 349 张 CID 谱图，1 349 张 ETD 谱图. 修饰类型包括乙酰化、磷酸化和甲基化(包括单甲基化、二甲基化和三甲基化). 这里我们仅比较时间和空间效率，三个软件的时间和空间开销如表 1.

Table 1　Performance comparison of pTop, MS-Align and ProSight

表 1　pTop、MS-Align+与 ProSight 性能比较

	pTop	MS-Align+	ProSight
时间	36 min	108 min	35 min
空间	35 M	4 G	267 M

pTop 和 MS-Align+采用 MS-Deconv 软件进行预处理，ProSight 采用 Thrash 进行预处理.

从时间和空间的开销上来看 pTop 优势明显，由于 ProSight 需要离线建立注释的数据库，在时间和空间上开销很大. MS-Align+考虑了蛋白质序列的任意位置是否发生修饰，而修饰的质量来自实验谱图中对应的谱峰与理论碎片离子偏移的质量，属于意外修饰. pTop 采用了新的技术策略排除掉了很多不可能的修饰组合，即排除掉了很多不可能的蛋白质变体. 例如，蛋白质 H4 上考虑乙酰化、磷酸化和甲基化等 5 种修饰，限制最多允许 10 个修饰位点时可能的蛋白质变体有上百亿种. 相比于 MS-Align+考虑上百亿种蛋白质变体，pTop 大大缩减了搜索空间，提高了搜索的时间和空间效率. 在鉴定精度方面，三个软件相当，由于当前发表的数据集还很少，在大规模的数据集上进行评估还无法完成，当前 pTop 的精度还有较大的提升空间.

5　总结与展望

基于完整蛋白质的质谱分析(top-down)与基于酶切肽段的质谱分析(bottom-up)构成了互补的蛋白质组技术体系，这两种分析方法各有优势和适用的问题，通过总结前文，我们将它们的对比列入表 2 中.

Table 2　Comparison of the techniques for bottom-up and top-down proteomics

表 2　Bottom-up 与 Top-down 蛋白质组技术方法的比较

比较内容	Bottom-up	Top-down
酶切	常用胰蛋白酶等	不经过酶切
分析对象	酶切后的肽段	完整蛋白质
分离技术	相对比较成熟、多维液相色谱使用广泛、通量高	多维分离、通量较低、自动化和分离效率有待提高
质谱技术	灵敏度高、谱图复杂度较低、低电荷谱峰典型	灵敏度较低、谱图复杂度高、高电荷谱峰比例高
鉴定软件	可选软件较多、应用广泛	实用软件很少，常需要手工处理
翻译后修饰	基于肽段上的修饰鉴定，一般容许最大 3 个修饰点	完整蛋白质上可能的修饰位点增多，如 10～20 个修饰位点
定量	多种定量方法、应该广泛	研究很少
规模化	可实现大规模蛋白质鉴定	刻画蛋白质的多翻译后修饰

虽然 TD 蛋白质组学在完整蛋白质的分离技术与质谱技术上发展很快，已经实现了高通量的蛋白质分离与鉴定，但是整体上仍然存在很多技术上的困难，例如对于不同人类细胞类型的蛋白质规模化鉴定、蛋白质变体的定量问题和高效的数据处理软件等等. 如果人的每个细胞类型预计平均有 25 万个蛋白质变体，考虑 4 000 种不同的细胞类型，那么在基于细胞的人类蛋白质组计划中就需要鉴定 10 亿个蛋白质变体，目前国际上正在执行中的该项计划期望能在 2030 年前完成[11]. 随着质谱技术的进一步发展，平均每个蛋白质变体的鉴定成本也会降低下来，图 8 给出了鉴定每个蛋白质变体的费用变化预测趋势.

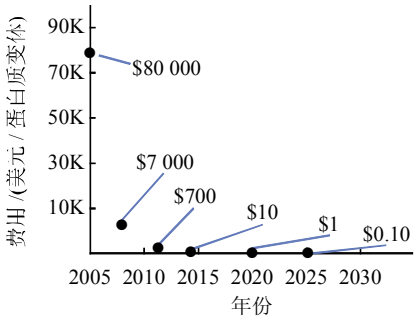

Fig. 8　Cost estimated for proteoform identification[11]

图 8　蛋白质变体的鉴定费用变化预测趋势[11]

除了大规模蛋白质变体的鉴定，未来的 TD蛋白质组学还将考虑蛋白质变体的定量问题．图 9给出了同一个蛋白质上发生了不同修饰情况的变体形式，分别采用 TD 和 BU 技术来分析它们对应的定量信息，从图 9 中可看出，完整的蛋白质分析技术发现蛋白质上同时发生甲基化和磷酸化的变体显著上调；而在 BU 技术分析中，只有发生了修饰的肽段显著上调，而没有获得修饰组合情况不同的变化趋势[35]．这样，在能够大量准确地鉴定蛋白质的变体后，对它们的定量就提到日程上来了，当前，在 TD 技术开发上还没有系统地研究变体定量的工作，未来这将成为重要的研究课题之一.

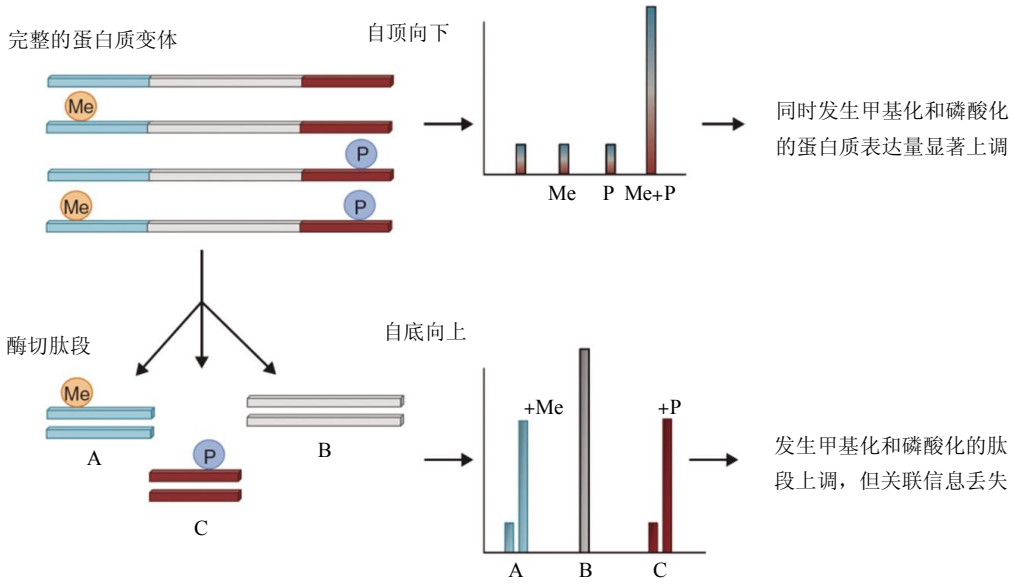

Fig. 9　Comparison between TD and BU for protein identification with multiple PTM[35]

图 9　TD 与 BU 分析发生多种翻译后修饰的蛋白质[35]

参　考　文　献

[1] 孙瑞祥, 付　岩, 李德泉, 等. 基于质谱技术的计算蛋白质组学研究. 中国科学 E 辑(信息科学), 2006, **36**(2): 222−234
Sun R X, Fu Y, Li D Q, *et al*. Science in China (E), 2006, **36**(2): 222−234

[2] Washburn M P, Wolters D, Yates III J R. Large scale analysis of the yeast proteome *via* multidimensional protein identification technology. Nature Biotechnology, 2001, **19**(3): 242−247

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

[4] Hebert A S, Richards A L, Bailey D J, *et al*. The one hour yeast proteome. Mol Cell Proteomics, 2014, **13**(1): 339−347

[5] Smith L M, Kelleher N L. Proteoform: a single term describing protein complexity. Nature Methods, 2013, **10**(3): 186−187

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

[7] Wu C, Tran J C, Zamdborg L, *et al*. A protease for 'middle-down' proteomics. Nature Methods, 2012, **9**(8): 822−824

[8] Xu P, Peng J. Characterization of polyubiquitin chain structure by middle-down mass spectrometry. Anal Chem, 2008, **80**(9): 3438−3444

[9] Fornelli L, Ayoub D, Aizikov K, *et al*. Middle-down analysis of monoclonal antibodies with electron transfer dissociation orbitrap Fourier transform mass spectrometry. Anal Chem, 2014, **86** (6): 3005−3012

[10] Kelleher N L. Top-down proteomics. Anal Chem, 2004, **76**(11), 196A-203A

[11] Kelleher N L. A cell-based approach to the human proteome project. J Am Soc Mass Spectrom, 2012, **23**(10): 1617-1624

[12] Chait B T. Mass spectrometry: bottom-up or top-down?. Science, 2006, **314**(5796): 65-66

[13] Siuti N, Kelleher N L. Decoding protein modifications using top-down mass spectrometry. Nature Methods, 2007, **4**(10): 817-821

[14] Arnaud C H. Top-down proteomics becomes reality. Chem Eng News, 2013, **91**(20): 11-17

[15] Whitelegge J. Intact protein mass spectrometry and top-down proteomics. Expert Rev Proteomics, 2013, **10**(2): 127-129

[16] Bogdanov B, Smith R D. Proteomics by FTICR mass spectrometry: top down and bottom up. Mass Spectrom Rev, 2005, **24**(2): 168-200

[17] Lane L, Bairoch A, Beavis R C, *et al*. Metrics for the human proteome project——2013-2014 and strategies for finding missing proteins. J Proteome Research, 2014, **13**(1): 15-20

[18] Hu Q, Noll R J, Li H, *et al*. The Orbitrap: a new mass spectrometer. J Mass Spectrom, 2005, **40**(4): 430-443

[19] Macek B, Waanders L F, Olsen J V, *et al*. Top-down protein sequencing and MS3 on a hybrid linear quadrupole ion trap-Orbitrap mass spectrometer. Mol Cell Proteomics, 2006, **5** (5): 949-958

[20] Ahlf D R, Compton P D, Tran J C, *et al*. Evaluation of the compact high-field Orbitrap for top-down proteomics of human cells. J Proteome Research, 2012, **11**(8): 4308-4314

[21] Zubarev R A, Makarov A. Orbitrap mass spectrometry. Anal Chem. 2013, **85**(11): 5288-5296

[22]Consortium for Top Down Proteomics, http: //www. topdownproteomics.org

[23] Kelleher N L, Ljiljana Paša-Tolić. 25th ASMS sanibel conference on top down mass spectrometry. J Am Soc Mass Spectrom, 2013, **24**(7): 983-985

[24] Top down proteomics virtual issue (ACS publication), http: //pubs. acs.org/page/vi/2013/topdown.html

[25] Top down presentations on 61st ASMS Conference on Mass Spectrometry and Allied Topics: http://topdownproteomics.org/news/conferences/item/61st-asms-conference-on-mass-spectrometry-and-allied-topics

[26] VerBerkmoes N C, Bundy J L, Hauser L, *et al*. Integrating "top-down" and "bottom-up" mass spectrometric approaches for proteomic analysis of shewanella oneidensis. J Proteome Research, 2002, **1**(3): 239-252

[27] Pesavento J J, Kim Y B, Taylor G K, *et al*. Shotgun annotation of histone modifications: a new approach for streamlined characterization of proteins by top down mass spectrometry. J Am Chem Soc, 2004, **126**(11): 3386-3387

[28] Peng Y, Chen X, Sato T, *et al*. Purification and high-resolution top-down mass spectrometric characterization of human salivary α-amylase. Anal Chem, 2012, **84**(7): 3339-3346

[29] Peng Y, Chen X, Zhang H, *et al*. Top-down targeted proteomics for deep sequencing of tropomyosin isoforms. J Proteome Research, 2013, **12**(1): 187-198

[30] Edwards R L, Griffiths P, Bunch J, *et al*. Top-down proteomics and direct surface sampling of neonatal dried blood spots: diagnosis of unknown hemoglobin variants. J Am Soc Mass Spectrom, 2012, **23**(11): 1921-1930

[31] Whitelegge J, Halgand F, Souda P, *et al*. Top-down mass spectrometry of integral membrane proteins. Expert Rev Proteomics, 2006, **3**(6): 585-596

[32] Catherman A D, Li M, Tran J C, *et al*. Top down proteomics of human membrane proteins from enriched mitochondrial fractions. Anal Chem, 2013, **85**(3): 1880-1888

[33] Ye H, Mandal R, Catherman A, *et al*. Top-down proteomics with mass spectrometry imaging: a pilot study towards discovery of biomarkers for neurodevelopmental disorders. PLoS One, 2014, **9**(4): e92831

[34] McLafferty F W, Breuker K, Jin M, *et al*. Top-down MS, a powerful complement to the high capabilities of proteolysis proteomics. FEBS J, 2007, **274**(1): 6256-6268

[35] Ahlf D R, Thomas P M, Kelleher N L. Developing top down proteomics to maximize proteome and sequence coverage from cells and tissues. Curr Opin Chem Biol, 2013, **17**(5): 789-794

[36] Savaryn J P, Catherman A D, Thomas P M, *et al*. The emergence of top-down proteomics in clinical research. Genome Medicine, 2013, **5**: 53

[37] Catherman A D, Skinner O S, Kelleher N L. Top down proteomics: facts and perspectives. Biochem Biophys Res Commun, 2014, **445**(4): 683-693

[38] Han X, Jin M, Breuker K, *et al*. Extending top-down mass spectrometry to proteins with masses greater than 200 kilodaltons. Science, 2006, **314**(5796): 109-112

[39] Li Y, Compton P D, Tran J C, *et al*. Optimizing capillary electrophoresis for top-down proteomics of 30-80 kDa proteins. Proteomics, 2014, **14**(10): 1158-1164

[40] Compton P D, Zamdborg L, Thomas P M, *et al*. On the scalability and requirements of whole protein mass spectrometry. Anal Chem, 2011, **83**(17): 6868-6874

[41] Tran J C, Zamdborg L, Ahlf D R. *et al*. Mapping intact protein isoforms in discovery mode using top-down proteomics. Nature, 2011, **480**(7376): 254-258

[42] Catherman A D, Durbin K R, Ahlf D R, *et al*. Large-scale top down proteomics of the human proteome: membrane proteins, mitochondria, and senescence. Mol Cell Proteomics, 2013, **12**(12): 3465-3473

[43] Tian Z, Zhao R, Tolic N, *et al*. Two-dimensional liquid chromatography system for online top-down mass spectrometry. Proteomics, 2010, **10**(20): 3610-3620

[44] Tian Z, Tolic N, Zhao R, *et al*. Enhanced top-down characterization of histone post-translational modifications. Genome Biology, 2012, **13**: R86

[45] Horn D M, Zubarev R A, McLafferty F W. Automated reduction

and interpretation of high resolution electrospray mass spectra of large molecules. J Am Soc Mass Spectrom, 2000, **11**(4): 320−332

[46] Zubarev R A, Kelleher N L, McLafferty F W. Electron capture dissociation of multiply charged protein cations. a nonergodic process. J Am Chem Soc, 1998, **120**(13): 3265−3266

[47] Zubarev R A, Horn D M, Fridriksson E K, *et al*. Electron capture dissociation for structural characterization of multiply charged protein cations, Anal Chem, 2000, **72**(3): 563−573

[48] Pan J, Borchers C H. Top-down structural analysis of post translationally modified proteins by Fourier transform ion cyclotron resonance-MS with hydrogen/deuterium exchange and electron capture dissociation. Proteomics, 2013, **13**(6): 974−981

[49] Syka J E P, Coon J J, Schroeder M J, *et al*. Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. Proc Natl Acad Sci USA, 2004, **101**(26): 9528−9533

[50] 孙瑞祥, 董梦秋, 迟 浩, 等. 基于电子捕获裂解 / 电子转运裂解串联质谱技术的蛋白质组学研究. 生物化学与生物物理进展, 2010, **37**(1): 94−102
Sun R X, Dong M Q, Chi H, *et al*. Prog Biochem Biophys, 2010, **37**(1): 94−102

[51] 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 Research, 2010, **9**(12): 6354−6367

[52] Chi A, Bai D L, Geer L Y, *et al*. Analysis of intact proteins on a chromatographic time scale by electron transfer dissociation tandem mass spectrometry. Int J Mass Spectrom, 2007, **259** (3): 197−203

[53] Drabik A, Bodzon-Kulakowska A, Suder P. Application of the ETD/PTR reactions in top-down proteomics as a faster alternative to bottom-up nano LC-MS/MS protein identification. J Mass Spectrom, 2012, **47**(10): 1347−1352

[54] Tsybin Y O, Fornelli L, Stoermer C, *et al*. Structural analysis of intact monoclonal antibodies by electron transfer dissociation mass spectrometry. Anal Chem, 2011, **83**(23): 8919−8927

[55] Bunger M K, Cargile B J, Ngunjiri A, *et al*. Automated proteomics of *E. coli via* top-down electron-transfer dissociation mass spectrometry. Anal Chem, 2008, **80**(5): 1459−1467

[56] Michalski A, Damoc E, Lange O, *et al*. Ultra high resolution linear ion trap Orbitrap mass spectrometer (Orbitrap Elite) facilitates top down LC MS/MS and versatile peptide fragmentation modes. Mol Cell Proteomics, 2012, **11**(3): O111.013698

[57] Senko M W, Remes P M, Canterbury J D, *et al*. Novel parallelized quadrupole/linear ion trap/Orbitrap tribrid mass spectrometer improving proteome coverage and peptide identification rates. Anal Chem, 2013, **85**(24): 11710−11714

[58] Zhou H, Ning Z, Starr A E, *et al*. Advancements in top-down proteomics. Anal Chem, 2012 , **84**(2): 720−734

[59] Liu X, Hengel S, Wu S, *et al*. Identification of ultramodified proteins using top-down spectra. RECOMB 2013, LNBI 7821, 132−144

[60] Liu X, Inbar Y, Dorrestein P C, *et al*. Deconvolution and database search of complex tandem mass spectra of intact proteins: a combinatorial approach. Mol Cell Proteomics, 2010, **9**(12): 2772−2782

[61] Meng F Y, Cargile B J, Miller L M, *et al*. Informatics and multiplexing of intact protein identification in bacteria and the archaea. Nat Biotechnol, 2001, **19**(10): 952−957

[62] Taylor G K, Kim Y B, Forbes A J, *et al*. Web and database software for identification of intact proteins using "top down" mass spectrometry. Anal Chem, 2003, **75**(16): 4081−4086

[63] LeDuc R D, Taylor G K, Kim Y B, *et al*. ProSight PTM: an integrated environment for protein identification and characterization by top-down mass spectrometry. Nucleic Acids Res, 2004, **32**(Suppl 2): W340−W345

[64] Zamdborg L, LeDuc R D, Glowacz K J, *et al*. ProSight PTM 2.0: improved protein identification and characterization for top down mass spectrometry. Nucleic Acids Res, 2007, **35**(Suppl 2): W701−W706

[65] Perkel J M. Tearing the top off 'top-down' proteomics. BioTechniques, 2012, **53**(2): 75−78

[66] Liu X, Sirotkin Y, Shen Y, *et al*. Protein identification using top-down spectra. Mol Cell Proteomics, 2012, M111.008524

[67] Frank A M, Pesavento J J, Mizzen C A, *et al*. Interpreting top-down mass spectra using spectral alignment. Anal Chem, 2008, **80** (7): 2499−2505

[68] Liu X, Hengel S, Wu S, *et al*. Identification of ultramodified proteins using top-down tandem mass spectra. J Proteome Research, 2013, **12**(12): 5830−5838

[69] Tsai Y S, Scherl A, Shaw J L, *et al*. Precursor ion independent algorithm for top- down shotgun proteomics. J Am Soc Mass Spectrom, 2009, **20**(11): 2154−2166

[70] Li L, Tian Z. Interpreting raw biological mass spectra using isotopic mass-to-charge ratio and envelope fingerprinting. Rapid Commun Mass Spectrom, 2013, **27**(11): 1267−1277

[71] Wiese H, Kuhlmann K, Wiese S, *et al*. Comparison of alternative MS/MS and bioinformatics approaches for confident phosphorylation site localization. J Proteome Research, 2014, **13**(2): 1128−1137

Top-down Proteomics: The Large-scale Proteoform Identification*

SUN Rui-Xiang**, LUO Lan, CHI Hao, LIU Chao, HE Si-Min

(*Key Labortory of Intelligent Information Processing of Chinese Academy of Sciences* (*CAS*), *Institute of Computing Technology, Chinese Academy of Sciences, Beijing* 100190, China; *University of Chinese Academy of Sciences, Beijing* 100049, China)

Abstract With the rapid advancement of the high resolution mass spectrometry, top-down proteomics becomes the reality. Proteome research on the intact protein level will provide more precise and more abundant biological information. For example, it can detect the relationship between the multiple post-translational modifications. Due to the genetic mutation, alternative splicing of RNA and various post-translational modifications, one gene may produce multiple protein forms, now called 'proteoforms'. Top-down proteomics will help identify the proteoforms. The three pillar technologies in top-down proteomics are separation, mass spectrometry and bioinformatics from the point of view on the entire proteins. This paper reviews these technologies and puts more emphases on the bioinformatics related topics, including the mass spectral preprocessing, the database searching algorithms and the localization of post-translational modifications.

Key words top-down proteomics, tandem mass spectrometry, bioinformatics, protein identification

DOI: 10.16476/j.pibb.2014.0078

* This work was supported by a grant from National Basic Research Program of China (2013CB911203, 2010CB912701).

**Corresponding author.

Tel: 86-10-62600822, E-mail: rxsun@ict.ac.cn

Received: March 19, 2014 Accepted: June 27, 2014