PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2007, 34(8): 791~799
www.pibb.ac.cn

蛋白质组学中新蛋白质鉴定的研究方法和策略 *

马 洁[1,2] 吴松锋[1,2] 朱云平[1,2]**
([1]军事医学科学院放射与辐射医学研究所，北京 100850；
[2]蛋白质组学国家重点实验室，北京蛋白质组研究中心，北京 102266)

摘要 当前，基于生物质谱进行蛋白质鉴定的技术已经成为蛋白质组学研究的支撑技术之一. 产生的数据主要使用数据库搜索的方法进行处理，这种方法的一大缺陷是不能鉴定数据库中未包含的蛋白质，因此如何充分利用质谱数据对蛋白质组研究的意义很大，而新蛋白质鉴定更是其中一个重要的内容. 新蛋白质鉴定是蛋白质鉴定的一个方面，新蛋白质的定义按照序列和功能的已知程度分为 3 个层次；以蛋白质鉴定的方法为基础，目前新蛋白质鉴定的方法可分为 *de novo* 测序和相似序列搜索结合的方法以及搜索 EST、基因组等核酸数据库的方法 2 大类；两者各有利弊.存在各自的问题和相应处理的策略. 不同的研究者可以根据具体目的应用和发展不同的鉴定方法，同时新蛋白质的鉴定也将随着蛋白质组学研究的发展而更加完善.

关键词 蛋白质组学，*de novo* 测序，数据库搜索，新蛋白质鉴定
学科分类号 Q51，Q657.6

在后基因组时代，蛋白质组学已成为研究的新热点之一. 正如人类基因组计划的提前实现得益于自动化的基因测序技术，生物质谱技术的应用也使蛋白质组学研究发生了质的飞跃. 随着质谱仪器、蛋白质和肽段的分离技术及数据分析工具的发展，不同组织、器官或细胞的蛋白质表达谱研究也得到了迅速的发展.

蛋白质组学最基本的目的就是定性和定量地鉴定一个细胞或组织中的全部蛋白质[1]. 随着蛋白质组学研究的深入，人们越来越关注其完整性，希望找到更多更新的蛋白质，于是充分挖掘质谱数据进行新蛋白质鉴定受到重视. 鉴定新蛋白质最基本的作用是保证蛋白质组学的完整性，并且可以完善和校正对基因组的注释，进一步发现新的基因编码区及新的基因模式[2]. 蛋白质序列信息允许对编码蛋白的基因组和转录组进行直接确认，鉴定新蛋白质则从实验上提供了新的转录产物的客观证据. 另外，鉴定到某物种、组织细胞某种特定状态下的新蛋白质对该组织器官新功能的认识，以及发现新的疾病诊断靶标具有重要的意义.

本文首先介绍了蛋白质鉴定的背景以及新蛋白质鉴定问题的产生，并从不同文献的研究观点中概括了新蛋白质的定义，然后从 *de novo* 测序和数据库搜索这 2 种最基本的蛋白质鉴定方法出发，详细讨论了目前鉴定新蛋白质的方法，最后分析了新蛋白质鉴定中存在的问题和相应的对策.

1 蛋白质鉴定的发展和新蛋白质鉴定问题的产生

传统的蛋白质鉴定基于 1950 年 Edman 提出的 Edman 降解法进行氨基酸测序，人们通过 Edman 降解法可以一个一个精确地确定氨基酸序列.但是该方法费时费力，难以实现高通量分析. 1980 年，Shimonishi 等[3]在将质谱和 Edman 降解法结合对肽段混合物进行测序，并针对该方法开发了相应的计算程序.

1984 年 Sakurai 等[4]基于全面搜索模式(穷举法)发展了自动化的从头(*de novo*)肽段测序算法，即产生所有符合母离子质量误差范围内的氨基酸序列及其组合的理论谱图，与实验谱图比较找出最佳匹配结果. 全局法会随着肽段序列长度的增加而产生“组合爆炸”问题. 之后，一些改进的算法也陆续出

*国家重点基础研究发展计划(973)(2006CB910803), 国家高技术研究发展计划(863) (2006AA02A312) 和创新研究群体科学基金(30621063)资助项目.
** 通讯联系人. Tel: 010-80727777-1223, E-mail: zhuyp@hupo.org.cn
收稿日期：2007-01-08，接受日期：2007-04-06

现，如局部法(图论法)把质谱峰转化成质谱图中的节点，通过构建最佳路径寻找氨基酸序列，大大提高了从头测序的效率[5]. 到 20 世纪 80 年代末，由于电喷雾电离(ESI)和基质辅助激光解吸电离(MADIL) 2 种“软电离”技术的出现，使得蛋白质等生物大分子的质谱鉴定成为可能. 20 世纪 90 年代以后公共可得的蛋白质序列逐渐增多，利用肽质量指纹(PMF)、串联质谱(MS/MS)数据和序列标签法，通过搜索已知蛋白质序列鉴定蛋白质的算法和软件相继出现，蛋白质组学研究得到飞跃发展.

随着国际人类基因组计划的完成，蛋白质序列数据库也逐渐完善，蛋白质组学研究从全局考虑希望耗费最少的时间和资源，鉴定到尽量多的蛋白质，而将串联质谱和数据库搜索结合来鉴定蛋白质，可以满足蛋白质组学研究高通量、自动化的要求，逐渐成为人类蛋白质组表达谱研究的重要方法之一. 一次批量质谱实验往往得到数以万计的谱图文件，常用的搜库软件 SEQUEST 和 Mascot 可以实现快速、有效的肽段鉴定. 但是数据库搜索方法也存在一些局限性：首先，该方法依赖于现有公共数据库以及数据库中蛋白质条目的完整性和正确性. 如果所研究物种的基因组没有被完全测序，或有部分蛋白质在数据库中没有对应序列，使用这种方法将无法做出正确的鉴定[6]. 另外，可能因为搜索引擎打分算法的不足而遗漏一些肽段 - 质谱图的匹配，导致不能鉴定或错误鉴定一些序列特异的肽碎片或肽侧链碎片，产生假阴性 / 假阳性结果. 而在另一些情况下，因为存在点突变、核酸多态性的存在，或者目标肽段含有未知的修饰等，由于搜库算法或参数设置中并没有考虑这些情况，同样也无法给出正确的匹配结果. 并且随着数据库中序列条目的增加、数据量的增大，数据库搜索方法的速度和效率会越来越低[7].

如上所述，虽然串联质谱和数据库搜索联用已成为蛋白质鉴定的主流方法，但当前的数据库搜索策略并不能完全解决蛋白质鉴定的问题. 蛋白质组学研究提出了进一步的要求：如何充分利用质谱数据，在保证高通量的同时提高蛋白质鉴定的完整性，并找到更多具有生物学功能意义的新蛋白质，新蛋白质鉴定的问题被提上议程并越来越受到重视. 实际上，新蛋白质鉴定只是蛋白质鉴定的一个方面，新蛋白质鉴定的问题一直存在于蛋白质组研究的发展过程之中，只是随着组学研究的逐渐成熟和人们要求的提高而凸现出来，并且它以蛋白质组学的研究方法为基础，随着组学研究的发展而发展. 人类基因组的测序结果提供了大量可能的编码蛋白，同时许多原核 / 真核模式生物的基因组测序也相继完成，使得人们使用原始的 MS/MS 搜索完整的、未经修饰的基因组数据库鉴定新蛋白质成为可能. 自从 Yates 等[8]在 1995 年第一次使用 MS/MS 搜索核酸序列之后，越来越多的研究者直接将 EST 或者核酸序列数据库翻译成相应的开放阅读框，对串联质谱数据进行最完整的鉴定，同时发现新蛋白质[9~12]. 另外，由于基因组测序本身的不完全性，还存在着一小部分空缺——基因组未测或测错的序列，也为新蛋白质鉴定提供了一定的发展空间.

2 新蛋白质的定义

虽然许多研究中都考虑了新蛋白质的鉴定问题，但关于新蛋白质的定义，不同的文献有不同说法，总体而言可以概括为以下 4 大类：

1. 发现具有新功能的已知蛋白质. 很多实验研究将发现具有新功能(包括定位、生理功能、生物标记物、相互作用等)的蛋白质称为新蛋白质(novel protein). Grønborg 等[13]使用 SILAC 技术对胰腺癌细胞和正常细胞进行差异蛋白质组学分析，在鉴定到的 145 个表达有差异的分泌性蛋白中，将一些在以前的研究中认为与胰腺癌没有关联的蛋白质称为新的胰腺癌标记蛋白. Ostrowski 等[14]在研究人类纤毛蛋白质组学时得到的新蛋白质，是指以往研究没有鉴定为纤毛组织成分的那些蛋白质，包括脑特异蛋白、精子表面蛋白以及视网膜色点蛋白等，以及一些匹配上预测 / 假设序列的蛋白质，并通过蛋白质印迹或免疫共沉淀等实验证明了其中一些蛋白质确实在纤毛或者有纤毛的细胞中表达，这些蛋白质可能为其新成分.

2. 蛋白质编码序列已被推测出来但未经验证、功能完全未知的蛋白质. Ram 等[15]将数据库中注释为“hypothetical”的基因编码的蛋白质认为是新蛋白质，这些蛋白质在数据库中有对应的基因编码序列而功能未知. Molina 等[16]认为鉴定到在数据库中注释为预测转录本的蛋白质代表新蛋白质，其中一个分泌型蛋白的功能完全未知. Kristiansen 等[17]对人类胆汁采用 RefSeq 数据库作基础鉴定，NCBI 的 nr 库作为补充鉴定，鉴定了一些新蛋白质包括功能未知的蛋白质.

3. 基因组数据库中包含、但未预测出相应编码蛋白的蛋白质. Smith 等[18]将基因组序列按 6 个阅读

框翻译成蛋白质序列来鉴定蛋白质，和 nr 数据库 BLAST 比对 E-value 没过阈值的 84 个蛋白质为在 nr 库中找不到相似序列的蛋白质，认为是基因编码的新蛋白质. Nesvizhskii 等[19]认为新肽段是指通过搜索未注释的基因组数据库鉴定的，但其序列在主要的蛋白质序列数据库中不存在的肽段. HPPP (人类血浆蛋白质组计划) 研究通过搜索基因组数据，获得了 118 个以往没有相应蛋白质报道的肽段序列，他们认为其中有些新蛋白质是因为基因注释错误或不完全所导致[2].

4. 序列完全未知的蛋白质. 这些蛋白质主要由于研究物种的基因组没有测序、测序不完全，或者基因预测软件错误导致所推测的开放阅读框(ORFs)错误，以及序列存在可变间接体、修饰等原因造成序列未知. Andersen 等[12]在拟南芥基因组测序不完全的情况下鉴定了 4 个新蛋白质，其结果被新公布较完整的基因组信息所确认. Matis 等[20]研究的菌类基因组尚未测序，他们通过 *de novo* 测序并和其他菌类的蛋白质数据库同源性比对，确定了一个可能具有脱氢酶活性的新蛋白质.

此外，某些数据库中也有自己定义新蛋白质的标准. 例如在 Ensemble 中将那些可以对应到 Swiss-Prot、RefSeq 或 SPTrEMBL 中的基因归为已知基因，而那些没有对应序列的称为新基因，由这些新基因得到的蛋白质就是新蛋白质.

大规模蛋白质表达谱研究中，鉴定新蛋白质很重要的一个目的是为了补充鸟枪蛋白质组学鉴定结果. 因此，本文主要讨论常用搜库软件和一般蛋白质数据库鉴定不到的、但在研究的组织或细胞中存在的那些蛋白质，并将它们统称为“新蛋白质”，并按照其序列的未知程度分为 3 个层次：第一个层次为已知部分功能，通过实验发现其新功能的新蛋白质；第二个层次为序列已知功能未知的蛋白质——这些新蛋白质常在搜库选用的数据库中没有对应序列，但其序列出现在更完备的蛋白质或基因组数据库中，一般功能都未知；第三个层次为新蛋白质中最高层次，指序列完全未知的蛋白质.

3　鉴定新蛋白质的方法

通过串联质谱鉴定蛋白质有 2 种最基本的方法[21]：一是通过数据库搜索的方法将质谱与蛋白质序列或核苷酸序列相关联进行鉴定，也就是常说的肽碎片指纹法(PFF)；另一种方法是不依赖于任何序列数据库而直接由谱图得到肽段序列，即 *de novo* 测序，推断得到的完整或部分的肽序列通过搜索引擎与理论序列比较得到最终结果. 新蛋白质鉴定是以蛋白质鉴定为基础的，基于上述蛋白质鉴定方法可将新蛋白质鉴定的方法主要分为 2 大类：

3.1　*De novo* 测序和数据库搜索相结合

De novo 测序是一种直接解释串联质谱数据的方法. 每个串联质谱峰对应一对肽键的裂解，通过确定质谱中峰的离子类型可以得到这个峰对应肽段的前缀或后缀子序列的质量. 理论上，可以根据串联质谱中信号强度高的碎片离子之间的质量差和母离子信号确定离子类型，得到相应的 b 离子、y 离子或其他离子峰. 然后通过计算相邻的同类型碎片离子之间的质量差获得肽段的全长序列[22].

De novo 测序通过串联谱中的信息重建肽段序列，不受已知蛋白质或基因组数据库所包含信息的影响，在下面几个情况下体现了它在鉴定新蛋白质领域中的灵活性：a. 基因组测序不完全而使得数据库中未包含待鉴定的蛋白质序列；b. 基因测序和预测软件中的错误，导致预测的蛋白质数据库中没有对应序列或对应蛋白质序列不正确；c. 基因的可变剪接体翻译得到的新蛋白质序列、编码区单核苷酸多态性引起的不同蛋白质变体，以及含修饰信息的序列等，这样的序列在数据库中都没有对应条目. 在上述情况下数据库搜索方法往往无能为力，而 *de novo* 测序则提供了一种可能的方法.

实际应用中 *de novo* 测序的通量极低，并且得到的肽段序列仍需通过序列比对找到与其功能或序列同源的蛋白质才有意义. 一般的谱图通过 *de novo* 测序只能得到少数可靠性较高的短片段，与后续其他手段结合才能达到鉴定蛋白质的目的. 因此，常将 *de novo* 测序和序列比对相结合来鉴定新蛋白质. 按照产生短序列标签方法的不同具体分 2 种：

a. 应用同位素标记得到序列标签.

在样品预处理过程中加入诸如 ^{13}C、^{18}O 和 ^{15}N 等同位素标记，例如分离后的蛋白质在含有 ^{18}O 环境中用胰岛素酶酶解，^{18}O 原子会选择性标记 C 端羟基集团[23]. 随后经 MS/MS 碎裂，通过手工或软件解读串联谱图，根据谱图上特定离子固定质量单位的偏移读出对应序列，组装成肽段序列标签再进行下一步的数据库搜索.

Matis 等[20]研究某菌类的蛋白质组时，该物种的基因组尚未测序. 将样品在 ^{16}O 和 ^{18}O 含量为 1∶1 的水中酶解进行同位素标记，分别经 2 种不同的 ESI-QTOF (QTOF： quadrupole time-of-flight mass

spectrometer，四极杆 - 飞行时间串联质谱仪)质谱仪测序，按照 ^{18}O 标记的规则手工解读谱图得到有同位素标记的肽序列标签，分别使用序列比对软件 MPsrch、FASTS 和 WU-BLAST2 对数据库做同源性搜索，3 种软件都鉴定了 4 种新蛋白质，并通过功能研究和序列同源性分析最后确定了一个可以解释其脱氢酶活性的新蛋白.

b. 由 *de novo* 测序软件得到序列标签

不依赖于实验的处理过程，应用 *de novo* 测序软件直接从常规的串联质谱图中读取肽段序列. 常见 *de novo* 测序软件的典型输出结果是一系列有序、无序的短肽段序列标签，或残基质量标签，利用这些短片段也就组成了序列标签.

常用产生肽序列标签的软件见表 1. 一般的 *de novo* 测序软件都可以通过网页形式免费访问，能满足大部分实验鉴定的需要，如果需要本地化大批量的对质谱数据进行肽序列标签的读取可以购买商业版本的软件如 Peaks，或者下载免费的代码如 Pepnovo 等.

Table 1 List of *de novo* sequencing alogrithms

表 1 *De novo* 测序软件列表

软件名称	网址	参考文献
Peaks	http://www.bioinformaticssolutions.com/products/peaks/	[24]
PepSeq	http://www.pepseq.com/	–
Lutefisk	http://www.hairyfatguy.com/Lutefisk	[25]
Pepnovo	http://peptide.ucsd.edu/pepnovo.py	[7]
GutenTag	http://fields.scripps.edu/GutenTag/index.html	[26]
SeqMS	http://www.protein.osaka-u.ac.jp/rcsfp/profiling/SeqMS.html	[27]
SHERENGA	---------	[28]

由 *de novo* 测序软件得到的短序列标签将通过数据库搜索算法与理论序列相关联，找到一致或相似的蛋白质序列. BLAST、FASTA、Shotgun[29]是序列同源性搜索最基本的几种算法，最初运用 *de novo* 测序得到的短肽标签进行下一步蛋白质鉴定分析的软件大多是基于它们的改进，像 MS-BLAST、FASTS/FASTF 和 MS-Shotgun，一般都没有考虑 *de novo* 测序的错误. 随后基于 FASTA 算法改良的 CIDentify 考虑了同位素峰和氨基酸质量无法区分的情况. DeNovoID 这个算法依赖于氨基酸序列的组成而对顺序没有要求，所以其用于查询的标签既可以有序也可以无序，还可以包括质量标签. 同源性搜索工具 SPIDER 考虑了同源突变和 *de novo* 测序的错误. GutenTag 软件整合了肽序列标签的产生和同源性比较搜库 2 部分，但没有考虑序列的同源突变和修饰. 主要序列同源性比较软件见表 2[30~39].

关于这类方法的应用，早在 1994 年，Mann 和 Wilm[30]就提出了用肽段序列标签鉴定蛋白质的方法. 他们将(碎片质量 m_1 – 部分序列 – 碎片质量 m_2) 这种形式的肽段序列标签提交给软件 PeptideSearch，既可以鉴定数据库中包含的蛋白质，还能鉴定数据库中没有的带有修饰的蛋白质或者可能的新蛋白质. Kim 等[40]在研究一种基因组未测序的土壤细菌时，从 Mascot 搜索细菌 nr 库没有匹配结果的 MS/MS 谱图中手工挑取质量较好的，使用 PepSeq 得到部分肽序列，通过与 Mascot 结果比较，再经 BLAST 同源性搜索得到高可信度的鉴定蛋白质以及一些新蛋白质. Lilla 等[41]也是运用 *de novo* 测序软件 PepSeq 和同源性比较程序 Blast-p 鉴定了一个新蛋白质. Kim 等[42]对他们研究中 PMF 和 PFF 搜索库不能鉴定的结果应用 Masslynx 软件经 *de novo* 测序得到部分肽段序列，然后通过 NCBI 上的 Blast 和 MS-BLAST 搜索得到接近全长的较精确的匹配结果. Williams 等[43]同样也应用 PepSeq 得到序列标签，然后通过 MS-Pattern 搜索 NCBI 的 nr 库鉴定蛋白质.

Table 2 List of sequence similarity search algorithms

表 2 肽序列标签搜库软件列表

软件名称	网址	参考文献
PeptideSearch	http://www.narrador.embl-heidelberg.de/GroupPages/Homepage.html	[30]
PepSea	http://www.unb.br/cbsp/paginiciais/pepseaseqtag.htm	–
MS-Seq	http://prospector.ucsf.edu/prospector/4.0.7/html/msseq.htm	[31]
MS-Pattern	http://prospector.ucsf.edu/prospector/4.0.7/html/mspattern.htm	[31]
FASTS/FASTF	http://www.ebi.ac.uk/fasta33/	[32]
MS-Blast	http://dove.embl-heidelberg.de/Blast2/msblast.html	[33]
MS-Shotgun	--------	[34]
GutenTag	http://fields.scripps.edu/GutenTag/index.html	[26]
OpenSea	--------	[35]
CIDentify	http://ftp.virginia.edu/pub/fasta/CIDentify/	[25]
DeNovoID	http://prometheus.brc.mcw.edu/denovoid/	[36]
MultiTag	--------	[37]
Popitam	http://www.expasy.org/tools/popitam/	[38]
SPIDER	http://bif.csd.uwo.ca/spider/	[39]

3.2 分步、多策略的数据库搜索：采用 ESTs、基因组数据库

De novo 测序不能取代数据库搜索方法在自动化 - 高通量蛋白质组学研究中的应用，同样在大规模蛋白质组学研究中进行新蛋白质的鉴定，数据库搜索方法也是不可缺少的. 但因为该方法强烈地依赖于数据库中所包含的蛋白质序列，即使质量很好的谱图，如果数据库中不存在相应的条目也无法得到鉴定结果. 因此，常规的蛋白质数据库不能实现鉴定新蛋白质的目的. 同时，数据库搜索的方法也不能发现序列信息未知的蛋白质，即无法鉴定层次最高的新蛋白质，但是通过分步、多策略的数据库搜索，如采用 ESTs、基因组数据库等更广泛全面的数据库[19, 44~46]，可以鉴定前 2 个层次的新蛋白质.

大部分蛋白质组学研究在进行基本的蛋白质鉴定的同时也利用实验质谱数据进行补充鉴定[17]和新蛋白质挖掘[47]. 其主要策略就是我们这里提到的分步、多策略的数据库搜索. 早在 2001 年，Andersen 等[12]就应用二级质谱数据搜索基因组数据库，结合一级谱图分析完善鉴定结果. 这样不但对蛋白质鉴定起到一定的补充作用，并能校正软件预测的基因组 ORFs，还能鉴定新蛋白质，文中就用该策略鉴定了一个人的新蛋白质. 2001 年，Creasy 等[9]提出了高通量并且有效的蛋白质鉴定，实验数据应该搜索一系列逐渐增大的数据库：首先，是一个小而质量高的蛋白质数据库如 Swiss-Prot 数据库，没有匹配结果的数据再搜索一个全面的 EST 数据库，搜索基因组数据库是最后一步. Smith 等[18]在 2005 年根据纤毛虫的核酸遗传编码规则，直接将嗜热四膜虫的基因组序列分别按 6 个相位翻译成对应的氨基酸序列作为 Mascot 搜索的数据库，鉴定了 223 个高可信度的蛋白质，然后将鉴定结果与 NCBI 的 nr 数据库进行 BLAST 同源性比对，确定了 84 个 nr 库中没有相似蛋白质的结果，认为可能是新蛋白质. 到 2006 年，Nesvizhskii 等[19]使用了多种搜索库软件如 SEQUEST、Mascot 及 X!Tandem，以及不同的参数设置如改变质量误差范围、增加修饰等，以得到尽量多的鉴定结果，并进一步对质谱图按照质量分等级，在基本搜索库没有结果的谱图中挑取质量较好的再搜索基因组数据库进行新蛋白质挖掘. Ying 等[46]也将分步搜索策略应用于人类胎肝蛋白质组学研究的补充鉴定和新蛋白质挖掘中，在进行基本蛋白质鉴定的基础上，应用 Q-TOF 数据先后搜索了整合的蛋白质数据库、人类 EST 数据库和基因组数据，找到了更多的蛋白质包括一些新蛋白质.

在数据库搜索策略中采用基因组数据解释质谱数据，并进一步将鉴定结果与基因组综合比较，可以直观地进行转录组和蛋白质组的比较，并实现蛋白质组研究对基因组的补充. Desiere 等[48]利用 Ensembl 的分布式注释系统(Distributed Annotation System：DAS)，将 SEQUEST 搜库鉴定得到的肽段用 BLAST 对应到人类基因组上，并实现可视化. Kalume 等[32]利用质谱数据经 Mascot 搜 nr 和基因组数据库，再利用 DAS 实现可视化把鉴定的肽段匹配回基因组上，并通过实验蛋白质组学的鉴定结果

与基因组 TBLASTN 比对，确定已知的转录本、新转录本以及一个新基因，并且校正了一些已知转录本. HPPP 在人类血浆蛋白表达谱研究中也对血浆数据进行了新蛋白质鉴定的数据挖掘[2]，他们把人类全基因组按 6 种相位开放阅读框翻译成无遗漏的蛋白质数据库作为 X!Tandem 软件搜索的数据库，来鉴定血浆蛋白的串联质谱数据，并采用层层严格的质量控制标准来降低假阳性率，将质谱结果匹配到基因组序列中发现了新基因模式，以及可能的新基因编码区和新可变剪接体.

综上所述，应用完整的、未经注释的基因组序列来进行蛋白质的鉴定可以极大程度地挖掘质谱数据. 合理安排分步、多策略的数据库搜索，并使用合适的搜索引擎和数据库，可以实现批量、自动化的新蛋白质鉴定，是目前高通量蛋白质组学研究中补充鉴定新蛋白质很好的一种选择.

4 新蛋白质鉴定的问题和策略

虽然现在大部分蛋白质组学研究中都考虑了新蛋白质鉴定，但至今还没有完全形成比较系统的研究方法，新蛋白质鉴定中还存在很多无法回避的问题，不同的研究者会根据各自的实际情况选择相应的研究策略.

4.1 基于 *de novo* 测序方法鉴定新蛋白质的问题和对策

De novo 测序方法最大的优点是能找到数据库中不包含的序列，具有较大的灵活性. 近年来发展了许多用于 *de novo* 测序的算法，但总体上都对质谱图的质量要求比较高，如需使用 Q-TOF、LTQ-FT (linear ion trap-fourier transform ion cyclotron mass spectrometry，FT-ICR MS，线性离子阱 - 傅立叶变换离子回旋共振质谱仪) 等质谱仪产出的数据，一般的 *de novo* 测序算法的性能会随质谱数据质量的提高而提高. 使用短肽序列标签有较好的特异性，得到结果的可靠性较高，并可以配合一些实验方法更精确地鉴定新蛋白质. 如 Qin 等[49]对野兔的蛋白质测序时，数据库中包含很少的兔类蛋白质，他们应用同位素标记 LC-MS/MS 数据经 *de novo* 测序来鉴定野兔的新蛋白质，并搜索人的蛋白质和 ESTs 数据库来寻找同源的蛋白质. Uttenweiler-Joseph 等[50]也应用同位素标记和微分扫描技术得到的串联质谱文件，经 *de novo* 测序研究牛脑细胞蛋白质. 一般情况下，*de novo* 测序方法很难直接得到完整的肽段序列，得到的短氨基酸序列中间仍会包含一定概率的错误测序，而像 OpenSea、SPIDER 等有一定容错性的相似性序列搜索软件的发展，将在后续的同源性比对中缓解这一问题，有利于找到正确的新蛋白质序列. 总体上来说，用 *de novo* 测序结合后续的序列搜索鉴定新蛋白质比较适合小规模的研究，若能巧妙地与实验手段结合，并辅以一定的专家手工确认，将可能鉴定到置信度较高的新蛋白质.

4.2 基于数据库搜索方法鉴定新蛋白质的问题和对策

基于数据库搜索方法鉴定新蛋白质可以实现高通量，能够满足现在规模蛋白质组学研究的需要，但在构建搜索数据库、搜索引擎效率和对新蛋白质可信度的确认这 3 个方面仍然存在许多未解决的问题.

首先，用数据库搜索的方法鉴定蛋白质就必须考虑选用数据库的问题. 虽然选用完整的基因组或 EST 数据库理论上可以包括全部可能编码的蛋白质序列，但不是所有的搜索库软件都能直接搜索核酸序列，即使像 Mascot 可以直接搜索核酸序列，也不能处理过大的基因序列，所以 Kalume 等[32]将 FASTA 格式的基因组序列整理成 100 kb 大小进行搜库. 一般需要将核酸序列转换成氨基酸序列. Smith等[18]和 Fermin 等[2]直接将基因组序列按照 6 个阅读框翻译成氨基酸序列分别作为 Mascot 和 X!Tandem 的搜索数据库，但这样得到的序列并不代表实际意义上基因编码的蛋白质，往往包含过多的短肽段而且冗余性也较高，文献[2]就指出他们如此翻译得到的 ORFs 平均长度是 25.5 个氨基酸残基，而 3.14 版本 IPI 数据库中平均蛋白质长度是 438.5 个氨基酸，差别较大. Nesvizhskii 等[19]搜索人类 EST 序列数据库来鉴定新肽段以及含序列多态性的肽段时，为了减少数据库含有测序错误或者冗余的条目数，对数据库进行再筛选：要求所包含序列中出现的开放阅读框至少含有 50 个明确的氨基酸，并且该序列在数据库中出现至少 2 次.

其次，随着数据库序列条目的增加，搜索空间增大，如何发挥搜索引擎的效率，控制搜索时间也是不可忽视的问题. 对基因组数据库搜索大量的 MS/MS 谱图计算强度大，并且需要耗费很大的资源和时间，如果使用高质量的谱图不但可以最小化错误的鉴定，还能减少耗费在解释错误结果上的时间[19, 45]. 另外，上面提到的对数据库过滤更重要的一个作用是使数据库结构紧凑，与搜索原始的 EST

数据库相比大大减少了数据库搜索的时间[19].

再次，很重要的一点是如何控制鉴定到的新蛋白质的置信度，如果不能确定找到的新蛋白质是否真实存在，鉴定结果将毫无意义. 基因组或 EST 数据库远大于蛋白质数据库，并且含有一定的测序误差. 同时，因为数据库大发生随机匹配的概率也增大，并且因为不正确的预测开放阅读框和许多低质量的 EST 序列，以及低质量的串联质谱数据，将导致很多的错误鉴定[45]. 一般的研究对搜库软件进行简单卡值来区分正确和错误结果，像 Kalume 等[32]只取 Mascot 的结果中分值大于 30 分的并且要求包含一个有 4 个连续氨基酸的序列标签. Smith 等[18]对 Mascot 搜索基因组数据库得到的结果卡 25 分，之后再通过搜索 EST 和一个初步的基因预测来进一步确认. Fermin 等[2]在 HPPP 研究中比较重视控制结果假阳性率，他们利用泊松模型计算发生错误匹配的概率，通过 X!Tandem 分值和假阴性率的 ROC 曲线确定卡值，并且要求鉴定的肽段满足对应的质谱图只匹配一个肽段、并且该肽段序列包含在同一个 ORF 中等多种条件，很多结果还需要进一步的实验验证.

串联质谱鉴定蛋白质在蛋白质组学研究中的应用已经比较成熟，我们认为基于串联质谱数据进行新蛋白质的鉴定过程中，以后将更加重视对质谱图过滤和后续搜库结果的质量控制这些重要环节的研究. 如果能对常规鉴定中没有匹配的谱图首先进行过滤，挑选其中质量较好的谱图，可以大大提高后续搜索更广泛的 EST、基因组数据库的速度，并能减少产生假阳性匹配的概率. 再加上对结果严格而合理的质量控制，可以保证得到新蛋白质的可信度. 此外 *de novo* 测序软件的进一步完善和发展将鉴定到更多序列信息未知的新蛋白质，在完整基因组数据库中仍没有匹配结果的高质量谱图应该通过 *de novo* 测序来寻找可能的序列信息. 如果能挑取鉴定结果中部分高可信度的新蛋白质进行后续实验，则可以从实验的角度验证新蛋白质的存在.

5 结　　语

在蛋白质组学研究中，利用质谱技术鉴定蛋白质的方法和转录组研究中 EST 测序、SAGE 等技术相似，属于一种相对开放型的方法[51]. 获得质谱数据后，采用 *de novo* 方法推导鉴定蛋白质则是一种完全开放(open)的方法，不依赖于样品中的蛋白质是否已知，理论上任何样品都可能推导出相应的蛋白质序列；若采用数据库搜索的方法则属于一种封闭(close)的方法，只能鉴定数据库中存在的蛋白质. 而目前大规模高通量蛋白质组学已经并且还在产生的海量数据，需要自动化的数据分析方法来集中鉴定高置信度的蛋白质，不断发展的搜库软件使得数据库搜索策略基本能够满足蛋白质组学的要求，成为处理质谱数据最常用的方法. 但由于并不存在一个理想的数据库能保证完备的蛋白质鉴定，随着研究的发展，该方法的封闭性越来越不可忽视，鉴定新蛋白质在所难免.

本文首先回顾了蛋白质鉴定背景和新蛋白质鉴定问题的产生，从序列未知程度的 3 个层次总结了新蛋白质的定义，然后从 *de novo* 测序和数据库搜索 2 种鉴定蛋白质主要方法的角度总结了目前进行新蛋白质鉴定的方法，这 2 大类方法各有优劣，同时也存在各自的问题和已有研究的部分解决方式，但是目前新蛋白质鉴定并没有提出一种通用的方法. 一般的，不同的研究者会根据不同的研究目的选择最适合的策略：如果是大规模的蛋白质组学研究，需要筛选大量的数据，以数据库搜索方法为基础来鉴定新蛋白质应该是第一选择；如果是针对特定组织、器官或是寻找特定功能类型的蛋白质，辅以适当的实验手段更容易鉴定到特异性较强的新蛋白质；而 *de novo* 测序则适合对基因组未完全测序的物种进行新蛋白质鉴定. 本实验室在对当前数据库的调研、论证的基础上，已经针对规模蛋白质组表达谱的研究提出并应用了分步搜索策略[46, 52]，希望将新蛋白质的鉴定整合到蛋白质鉴定的过程中，发现整合的多数据库、分步搜索策略，并结合 *de novo* 测序方法是高通量蛋白质组学串联质谱数据鉴定新蛋白质的一种很好的选择. 展望未来，新蛋白质的鉴定将随着蛋白质组学研究的发展而进一步完善.

参 考 文 献

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

2 Fermin D, Allen B B, Thomas W, *et al*. Novel gene and gene model detection using a whole genome open reading frame analysis in proteomics. Genome Biology, 2006, **7** (4): R35

3 Shimonishi Y, Hong Y M, Kitagishi T, *et al*. Sequencing of peptide mixtures by edman degradation and field-desorption mass spectrometry. Euro J Biochem, 1980, **112** (2): 251～264

4 Sakurai T, Matsuo T, Matsuda H, *et al*. Paas 3: A computer program to determine probable sequence of peptides from mass spectrometric data. Biomedical Mass Spectrometry, 1984, **11** (8): 396～399

5 Bartels C. Fast algorithm for peptide sequencing by mass spectroscopy. Biomedical and Environmental Mass Spectrometry, 1990, **19** (6): 363～368

6 Leipzig J, Pevzner P, Heber S. The alternative splicing gallery (Asg): bridging the gap between genome and transcriptome. Nucleic Acids Research, 2004, **32** (13): 3977～3983

7 Frank A, Pevzner P. Pepnovo: *De novo* peptide sequencing *via* probabilistic network modeling. Analytical Chemistry, 2005, **77** (4): 964～973

8 Yates III J R, Eng J K, McCormack A L. Mining genomes: correlating tandem mass spectra of modified and unmodified peptides to sequences in nucleotide databases. Analytical Chemistry, 1995, **67** (18): 3202～3210

9 Creasy D M, Cottrell J S. Interrogating the human genome using uninterpreted mass spectrometry data. Proteomics, 2001, **1** (5): 651～667

10 Link A J, Eng J, Schieltz D M, *et al*. Direct analysis of protein complexes using mass spectrometry. Nature Biotechnology, 1999, **17** (7): 676～682

11 Neubauer G, King A, Rappsilber J, *et al*. Mass spectrometry and Est-database searching allows characterization of the multi-protein spliceosome complex. Nature Genetics, 1998, **20** (1): 46～50

12 Andersen J S, Mann M. Mass spectrometry allows direct identification of proteins in large genomes. Proteomics, 2001, **1** (5): 641～650

13 Grønborg M, Kristiansen T Z, Iwahori A, *et al*. Biomarker discovery from pancreatic cancer secretome using a differential proteomic approach. Molecular & Cellular Proteomics, 2006, **5** (1): 157～171

14 Ostrowski L E, Blackburn K, Radde K M, *et al*. A proteomic analysis of human ciliaidentification of novel components. Molecular & Cellular Proteomics, 2002, **1** (6): 451～465

15 Ram R J, VerBerkmoes N C, Thelen M P, *et al*. Community proteomics of a natural microbial biofilm. Science, 2005, **308** (5730): 1915～1920

16 Molina H, Bunkenborg J, Reddy G H, *et al*. A proteomic analysis of human hemodialysis fluid. Molecular & Cellular Proteomics, 2005, **4** (5): 637～650

17 Kristiansen T Z, Bunkenborg J, Gronborg M, *et al*. A proteomic analysis of human bile. Molecular & Cellular Proteomics, 2004, **3** (7): 715～728

18 Smith J C, Northey J G B, Garg J, *et al*. Robust method for proteome analysis by Ms/Ms using an entire translated genome: demonstration on the ciliome of tetrahymena thermophila. J Proteome Research, 2005, **4** (3): 909～919

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

20 Matis M, Akelj-Mavri M, Peter-Katalini J. Mass spectrometry and database search in the analysis of proteins from the fungus pleurotus ostreatus. Proteomics, 2005, **5** (1): 67～75

21 Lu B, Chen T. Algorithms for *de novo* peptide sequencing using tandem mass spectrometry. Drug Discovery Today, 2004, **2** (2): 85～90

22 王中胜, 朱云平, 贺福初. 肽序列从头测序算法. 军事医学科学院院刊, 2006, **30** (5): 465～467
Wang Z S, Zhu Y P, He F C. Bulletin of the Academy of Military Medical Sciences, 2006, **30** (5): 465～467

23 Shevchenko A, Chernushevich I, Ens W, *et al*. Rapid *de novo* peptide sequencing by a combination of nanoelectrospray, isotopic labeling and a quadrupole/time-of-flight mass spectrometer. Rapid Communications in Mass Spectrometry, 1997, **11** (9): 1015～1024

24 Ma B, Zhang K, 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

25 Taylor J A, Johnson R S. Sequence database searches via *de novo* peptide sequencing by tandem mass spectrometry. Rapid Communications in Mass Spectrometry, 1997, **11** (9): 1067～1075

26 Tabb D L, Saraf A, Yates 3rd J R. Gutentag: High-throughput sequence tagging via an empirically derived fragmentation model. Analytical Chemistry, 2003, **75** (23): 6415～6421

27 Fernandez-de-Cossio J, Gonzalez J, Besada V. A computer program to aid the sequencing of peptides in collision-activated decomposition experiments. Bioinformatics, 1995, **11** (4): 427～434

28 Cik V, Addona T A, Clauser K R, *et al*. *De novo* peptide sequencing via tandem mass spectrometry. J Computational Biology, 1999, **6** (3～4): 327～342

29 Pegg S C, Babbitt P C. Shotgun: getting more from sequence similarity searches. Bioinformatics, 1999, **15** (9): 729～740

30 Mann M, Wilm M. Error-tolerant identification of peptides in sequence databases by peptide sequence tags. Analytical Chemistry, 1994, **66** (24): 4390～4399

31 Clauser K R, Baker P, Burlingame A L. Role of accurate mass measurement (±10 ppm) in protein identification strategies employing Ms or Ms/Ms and database searching. Analytical Chemistry, 1999, **71** (14): 2871～2882

32 Kalume D E, Peri S, Reddy R, *et al*. Genome annotation of anopheles gambiae using mass spectrometry-derived data. BMC Genomics, 2005, 6: 128

33 Shevchenko A, Sunyaev S, Loboda A, *et al*. Charting the proteomes of organisms with unsequenced genomes by maldi-quadrupole time-of-flight mass spectrometry and Blast homology searching. Analytical Chemistry, 2001, **73** (9): 1917～1926

34 Huang L, Jacob R J, Pegg S C H, *et al*. Functional assignment of the 20 S proteasome from trypanosoma brucei using mass spectrometry and new bioinformatics approaches. J Biol Chem, 2001, **276** (30): 28327～28339

35 Searle B C, Dasari S, Turner M, *et al*. High-throughput identification of proteins and unanticipated sequence modifications using a mass-based alignment algorithm for Ms/Ms *de novo* sequencing results. Analytical Chemistry, 2004, **76** (8): 2220～2230

36 Halligan B D, Ruotti V, Twigger S N, *et al*. Denovoid: A Web-based tool for identifying peptides from sequence and mass tags deduced from *de novo* peptide sequencing by mass spectroscopy. Nucleic Acids Research, 2005, **33** (1): W376～W381

37 Sunyaev S, Liska A J, Golod A, *et al*. Multitag: multiple error-tolerant sequence tag search for the sequence-similarity

identification of proteins by mass spectrometry. Analytical Chemistry, 2003, **75** (6): 1307～1315

38 Geneva S. Popitam: towards new heuristic strategies to improve protein identification from tandem mass spectrometry data. Proteomics, 2003, **3** (6): 870～878

39 Han Y, Ma B, Zhang K. Spider: software for protein identification from sequence tags with *de novo* sequencing error. J Bioinformatics and Computational Biology, 2005, **3** (3): 697～716

40 Kim E A, Kim J Y, Kim S J, *et al*. Proteomic analysis of acinetobacter lwoffii K24 by 2-D gel electrophoresis and electrospray ionization quadrupole-time of flight mass spectrometry. J Microbiological Methods, 2004, **57** (3): 337～349

41 Lilla S, Pereira R, Hyslop S, *et al*. Purification and initial characterization of a novel protein with factor Xa activity from lonomia obliqua caterpillar spicules. J Mass Spectrometry, 2005, **40** (3): 405～412

42 Kim H J, Lee D Y, Lee D H, *et al*. Strategic proteome analysis of candida magnoliae with an unsequenced genome. Proteomics, 2004, **4** (11): 3588～3599

43 Williams T L, Monday S R, Edelson-Mammel S, *et al*. A top-down proteomics approach for differentiating thermal resistant strains of enterobacter sakazakii. Proteomics, 2005, **5** (16): 4161～4169

44 Nesvizhskii A I, Aebersold R. Analysis, statistical validation and dissemination of large-scale proteomics datasets generated by tandem MS. Drug Discovery Today, 2004, **9** (4): 173～181

45 Nesvizhskii A I, Aebersold R. Interpretation of shotgun proteomic data the protein inference problem. Molecular & Cellular Proteomics, 2005, **4** (10): 1419～1440

46 Ying W, Jiang Y, Guo L, *et al*. A dataset of human fetal liver proteome identified by subcellular fractionation and multiple protein separation and identification technology. Molecular & Cellular Proteomics, 2006, **5** (9): 1703～1707

47 Omenn G S, David J, Adamski M, *et al*. Overview of the hupo plasma proteome project: results from the pilot phase with 35 collaborating laboratories and multiple analytical groups, generating a core dataset of 3020 proteins and a publicly-available database. Proteomics, 2005, **5** (13): 3226～3245

48 Desiere F, Deutsch E W, Nesvizhskii A I, *et al*. Integration with the human genome of peptide sequences obtained by high-throughput mass spectrometry. Genome Biology, 2004, **6** (1): R9

49 Qin J, Herring C J, Zhang X. *De novo* peptide sequencing in an ion trap mass spectrometer with ^{18}O labeling. Rapid Communications in Mass Spectrometry, 1998, **12** (5): 209～216

50 Uttenweiler-Joseph S, Neubauer G, Christoforidis S, *et al*. Automated *de novo* sequencing of proteins using the differential scanning technique. Proteomics, 2001, **1** (5): 668～682

51 Scheel J. Yellow pages to the transcriptome. Pharmacogenomics, 2002, **3** (6): 791～807

52 吴松锋, 朱云平, 贺福初. 人类蛋白质组表达谱蛋白质鉴定的分步搜素策略. 遗传, 2005, **27** (5): 687～693
Wu S F, Zhu Y P, He F C. Hereditas (Beijing), 2005, **27** (5): 687～693

Methods and Strategies of Novel Proteins Identification in Proteomics*

MA Jie[1,2], WU Song-Feng[1,2], ZHU Yun-Ping[1,2]**

([1] *Beijing Institute of Radiation Medicine, Beijing* 100850, China;
[2] *State Key Laboratory of Proteomics, Beijing Proteome Research Center, Beijing* 102206, China)

Abstract The combination of tandem spectrometry and database searching is one of the most popular technologies for protein identification. However, only those proteins in the searching database could be identified, and current database is far from completeness. So it is necessary to mining the MS/MS data comprehensively, in which novel protein identification is the most important one. The definition of novel protein could be divided into three levels according to their annotations of sequences and functions. As a part of protein identification, the main approaches used to identify novel protein are basing on the following two different ways: *de novo* sequencing combined with similarity search and searching against nucleotide acid databases such as EST or genome databases. Several mature or newly developed methods and techniques were summarized, and the problems and strategies discussed here would be helpful for the related researches.

Key words proteomics, *de novo* sequencing, database searching, novel protein identification

*This work was supported by grants from National Basic Research Program of China (2006CB910803), Hi-Tech Research and Development Program of China (2006AA02A312) and The National Natural Science Foundation of China (30621063).

**Corresponding author . Tel: 86-10-8072777-1223, E-mail: zhuyp@hupo.org.cn

Received: January 8, 2007 Accepted: April 6, 2007