PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2010, 37(12): 1278~1288
www.pibb.ac.cn

串联质谱图谱从头测序算法研究进展 *

孙汉昌[1)] 张纪阳[1)] 刘 辉[1)] 张 伟[1)] 徐长明[1)] 马海滨[1)] 朱云平[2)] 谢红卫[1)**]

([1)] 国防科学技术大学机电工程与自动化学院自动控制系，长沙 410073；
[2)] 军事医学科学院放射与辐射医学研究所，北京蛋白质组研究中心，蛋白质组学国家重点实验室，北京 102206)

摘要 近年来，基于质谱技术的高通量蛋白质组学研究发展迅速，利用串联质谱图谱鉴定蛋白质是其数据处理中一个基础而又重要的环节．由于不需要利用蛋白质序列数据库，从头测序方法能够分析新物种或者基因组未测序物种的串联质谱数据，具有数据库搜索方法不可替代的优势．简要介绍高通量串联质谱图谱从头测序问题及其研究现状．归纳出几种典型的计算策略并分析了各种策略的优缺点．总结常用的从头测序算法和软件，介绍算法评估的各种指标和常用评估数据集，概括各种算法的特点，展望未来研究可能的发展方向．

关键词 肽段，串联质谱，从头测序，算法研究，计算蛋白质组学
学科分类号 Q51，TP301 **DOI**: 10.3724/SP.J.1206.2010.00226

在生命科学研究领域中，蛋白质组学[1-3]是一个相对年轻但充满活力并且飞速发展的学科．蛋白质组学致力于通过高通量的实验，分析和刻画细胞、组织或有机体中在特定条件下表达的整套蛋白质．经过几年的快速发展，蛋白质组学的理念、技术方法和实验策略在临床医学、生物化学、病理学和疾病治疗等相关研究领域中的应用越来越多[4-8]．

质谱分析法是高通量蛋白质组学不可或缺的实验技术，已经成功应用到蛋白质组学研究的方方面面[4, 9]．基于质谱技术的蛋白质组学实验会产出海量的质谱数据[10]，对这些高通量数据进行计算分析和统计推断，进而挖掘和利用蕴含其中的生物信息是一项巨大的挑战，其中，蛋白质鉴定是最基础也是最核心的问题之一．

在蛋白质组学中，利用串联质谱分析鉴定肽段序列，然后再推断样品中包含的蛋白质，是常用的高通量分析策略，称之为鸟枪法，其基本流程如图 1 所示．其中，肽段是通过蛋白质酶切产生的，而串联质谱分析产生的二级图谱(MS/MS spectrum)中包含肽段的序列信息．从图 1 中可以看出，数据分析的基本任务是利用二级图谱确定样品中存在的蛋白质，也称为图谱解析，在计算策略上可以分为三类：数据库搜索方法、从头测序方法和基于肽段序列标签的方法．数据库搜索的基本思路是“模板匹配”，将实验得到的图谱与从数据库理论酶切产生的肽段的理论预测图谱进行比对，按照一定的打分规则鉴定出匹配最好的肽段[11]．从头测序方法不依赖现有的数据库，根据肽段有规律碎裂的特点，直接从图谱中推导出肽段的序列，能够分析新物种或者基因组未测序物种的串联质谱数据，具有数据库搜索方法不可替代的优势．肽段序列标签方法则是前两种方法的折衷，先从图谱中直接推导出肽段序列的局部标签，然后用推导出的肽段标签搜索数据库．这种方法主要考虑了图谱中包含肽段序列信息不完全和生物蛋白质序列的相似性等特点，在修饰、突变和跨物种搜索数据库中应用较多．

* 国家重点基础研究发展计划(973)(2006CB910803, 2006CB910706, 2010CB912700), 国家高技术研究发展计划(863)(2006AA02A312), 国家科技重大专项(2008ZX10002-016, 2009ZX09301-002)和蛋白质组学国家重点实验室课题(SKLP-Y200811)资助项目．
** 通讯联系人．
Tel: 0731-84576311, E-mail: xhwei65@nudt.edu.cn
收稿日期：2010-04-27，接受日期：2010-07-05

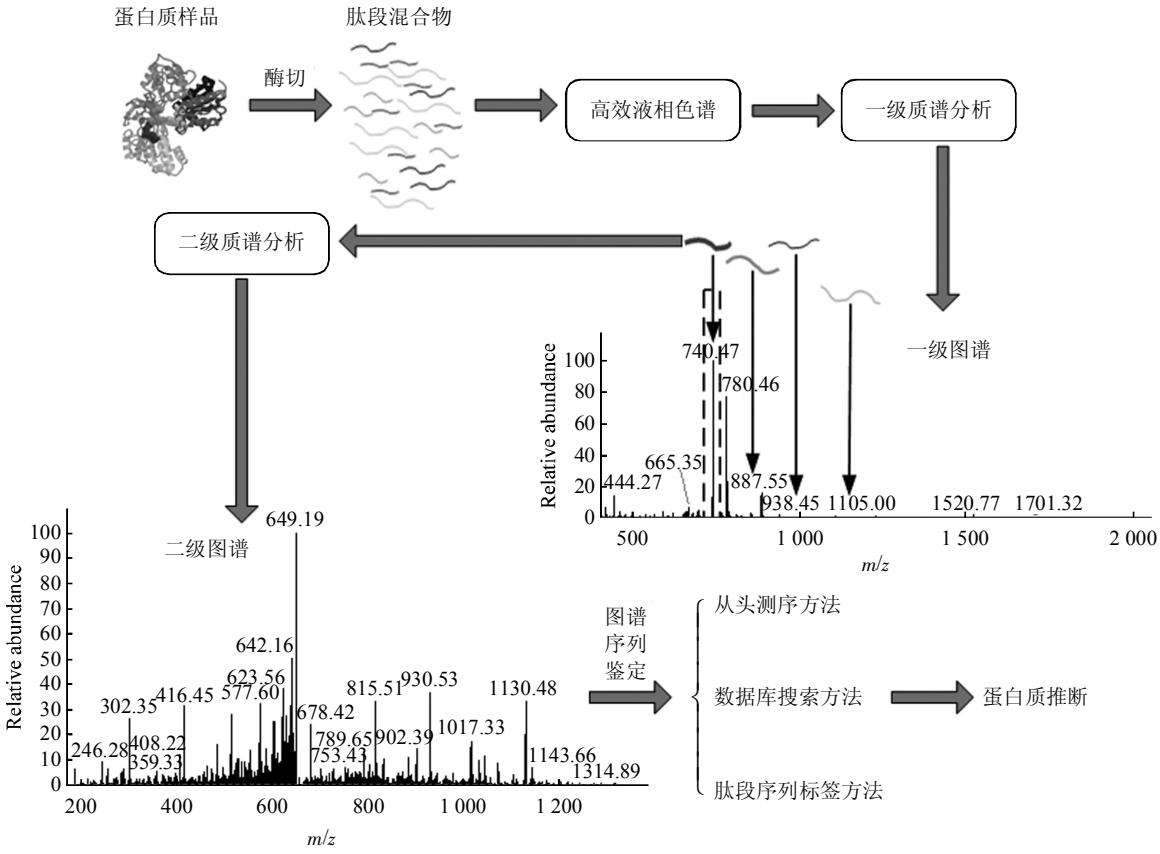

Fig. 1　Workflow of shotgun protein identification

图 1　使用鸟枪法鉴定蛋白质流程图

相对而言，在肽段序列鉴定的三类方法中，从头测序方法更加困难、更不成熟、因而有更多的问题有待进一步研究．本文将首先介绍串联质谱图谱从头测序问题，然后结合从头测序方法的研究现状，详细分析从头测序问题的几类典型解决途径，以及相关的算法评估指标和常用的数据集，最后给出从头测序算法未来可能的研究方向．

1　串联质谱图谱从头测序问题描述

在典型的高效液相色谱 - 串联质谱联用分析过程中，肽段混合物通过液相色谱分离后，经过电喷雾离子化，进入质谱进行一级、二级质谱分析，被选定的肽段离子在碰撞室与惰性气体作用而发生碰撞诱导解离．如图 2 所示，肽段主要有三种解离模式[12]，假定肽链含有 n 个氨基酸残基，肽段在第 m 和 m+1 个氨基酸残基之间发生解离，根据肽链断开的位置不同，会产生三类互补的离子，(a_m, x_{n-m})、(b_m, y_{n-m})和(c_m, z_{n-m})，其中，a 型离子、b 型离子和 c 型离子为 N 端离子，x 型离子、y 型离子和 z 型离子为 C 端离子，下标代表所含氨基酸残基的个数．在碰撞诱导解离(collision-induced dissociation, CID)图谱中，最常见的是 a 型离子、b 型离子和 y 型离子，其他类型离子较少出现．图 3 给出了双电荷母离子经过理论解离得到 b 型离子和 y 型离子的示意图，该肽段包含 7 个氨基酸残基，N 端的 b 型离子包含 3 个氨基酸残基，命名为 b_3，C 端的 y 型离子包含 4 个氨基酸残基，命名为 y_4.

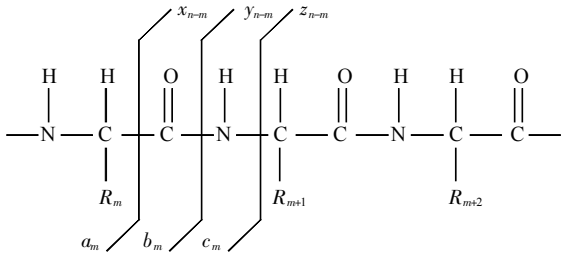

Fig. 2　Peptide fragmentation patterns and ion naming rules

图 2　肽段解离模式及离子命名规则

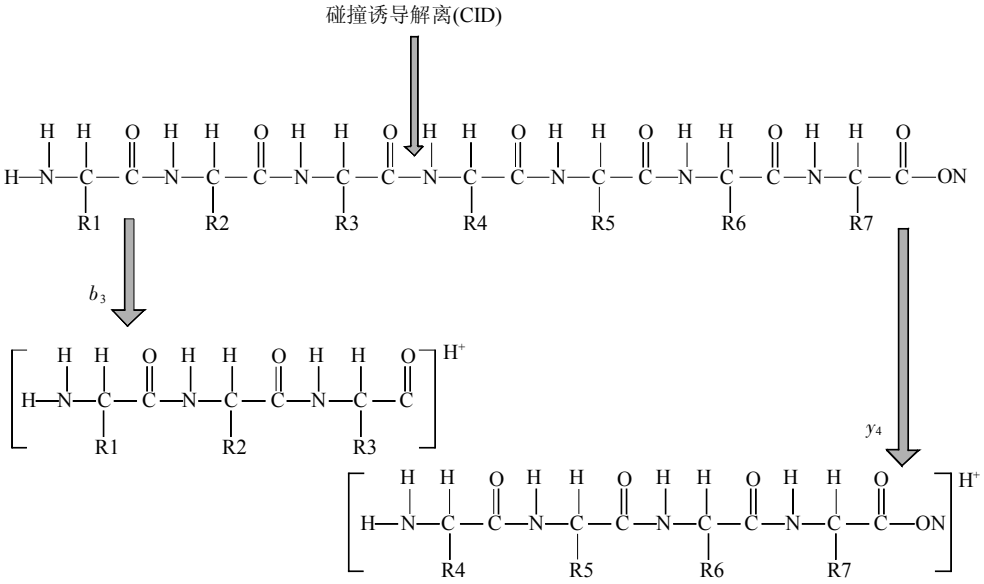

Fig. 3 A sample of collision induced dissociation
图 3 肽段理论解离示意图

理论上，串联质谱图谱从头测序问题可以描述为：已知母离子质量 M, 电荷 C, 母离子质量误差 ε, 碎片离子质量误差 δ，由实验得到的含有 k 个质谱峰(m_j, i_j)的图谱 S，其中，$1 \leqslant j \leqslant k$，$m_j$ 为峰对应的质量，i_j 为峰对应的强度，要求确定实验图谱 S 对应的真实肽段序列 P_r. 在多数情况下，由于实验图谱 S 中一些重要的碎片离子峰的丢失，要准确确定 P_r 将面临很大的困难，计算的时候，通常是先得到一组候选肽段$(P_1,P_2,P_3,\cdots\cdots P_{t\text{-}1},P_t)$，使得 $|M-M_{P_n}| < \varepsilon$，$1 \leqslant n \leqslant t$，$M_{P_n}$为候选肽段 P_n 对应的质量，然后再从 t 个候选肽段中找出与实验图谱 S 匹配最好的肽段 P_s.

串联质谱图谱从头测序的理论基础是：图谱中两个峰 Pk_i 和 Pk_j 对应的质量分别为 M_{Pk_i} 和 M_{Pk_j} $(1 \leqslant i, j \leqslant k)$，$k$ 为图谱中峰的个数，M_{AA_n} $(1 \leqslant n \leqslant 20)$为 20 种常见氨基酸残基对应的质量，$\lambda$ 为允许误差，如果上述参数满足 $\|M_{Pk_i}-M_{Pk_j}|-M_{AA_n}| < \lambda$，即认为 Pk_i 和 Pk_j 可能是母离子产生的碎片离子序列中相邻的两个离子生成的质谱峰，进而通过不断计算确定 b 离子系列$(b_1,b_2,b_3\cdots\cdots b_{l\text{-}1},b_l)$的位置($l$ 为真实肽段序列长度)或者 y 离子系列$(y_1,y_2,y_3\cdots\cdots y_{l\text{-}1},y_l)$的位置，最后通过得到的 b 离子或者 y 离子的位置推导出图谱对应的肽段序列信息.

2 串联质谱图谱从头测序计算策略

在实际的研究中，研究者们提出了各种各样的方法来解决串联质谱图谱从头测序问题. 总的来说，这些方法又可以分为三类，即穷举法、基于图论的方法以及综合方法. 下面分别归纳总结这三类方法.

2.1 穷举法

穷举法根据母离子质量列举出所有可能的候选肽段，然后将候选肽段和实验图谱做比对，找出最佳匹配候选肽段. 这种方法也可以认为是广义上的数据库搜索方法，搜库空间为 20^l，l 为肽段长度. 这类方法在 20 世纪 80 年代由 Sakurai 等[13]提出. 随着肽段长度或者母离子质量的增加，候选肽段的数量呈几何级数增长，例如，分子质量为 774 u 的肽段对应着 21 909 046 个可能的候选肽段[13]，基于穷举的方法只适用于长度较短的肽段，不适合推广. 另外一个不足是计算速度太慢，尽管有一些研究工作[14-18]加快了这类方法的计算速度，但因为它本身是一个 NP-hard 问题，加速计算并不能满足大规模质谱数据处理的需求. 2006 年，Olson 等[19]使用高精度的 MALDI TOF-TOF 仪器产出误差为 50mu 的高精度数据，能够处理母离子质量为

1 600 u 或者更多的图谱，但他们提出的方法对其他类型的仪器还不适用.

Lu 等[20]和 Xu 等[21]对穷举法进行了综述. 总的来说，基于穷举的方法能够鉴定较短的肽序列，但并不适用于较长的肽序列或低精度仪器产出的数据，其主要难点在于计算复杂度和极其相似的候选肽段序列的区分.

2.2　基于图论的方法

自从 Bartels[22]在 1990 年提出使用图论方法解决串联质谱图谱从头测序问题以来，许多研究者提出的算法都可以归于此类. 针对双电荷母离子，这类方法的基本流程可以概述如下：首先是图谱预处理过程，例如，去掉图谱中低丰度的峰，或者归并图谱中的同位素峰簇等；然后构建质谱峰连接图，即如果两个峰之间的质量差在误差范围内等于某个氨基酸残基的质量，就将这两个质谱峰作为两个顶点和一条边加入到(V, E)图中，质谱峰连接图构建完毕后，在(V, E)图中加入 b 型离子的起始点 1 和结束点 M–17 以及 y 型离子的起始点 19 和结束点 M+1，其中，M 为母离子质量，再在(V, E)图中搜索 b 型离子或者 y 型离子从起始点到结束点的路径，同时产生候选肽段；最后通过打分函数对候选肽段进行排序和输出. 其流程图如图 4 所示.

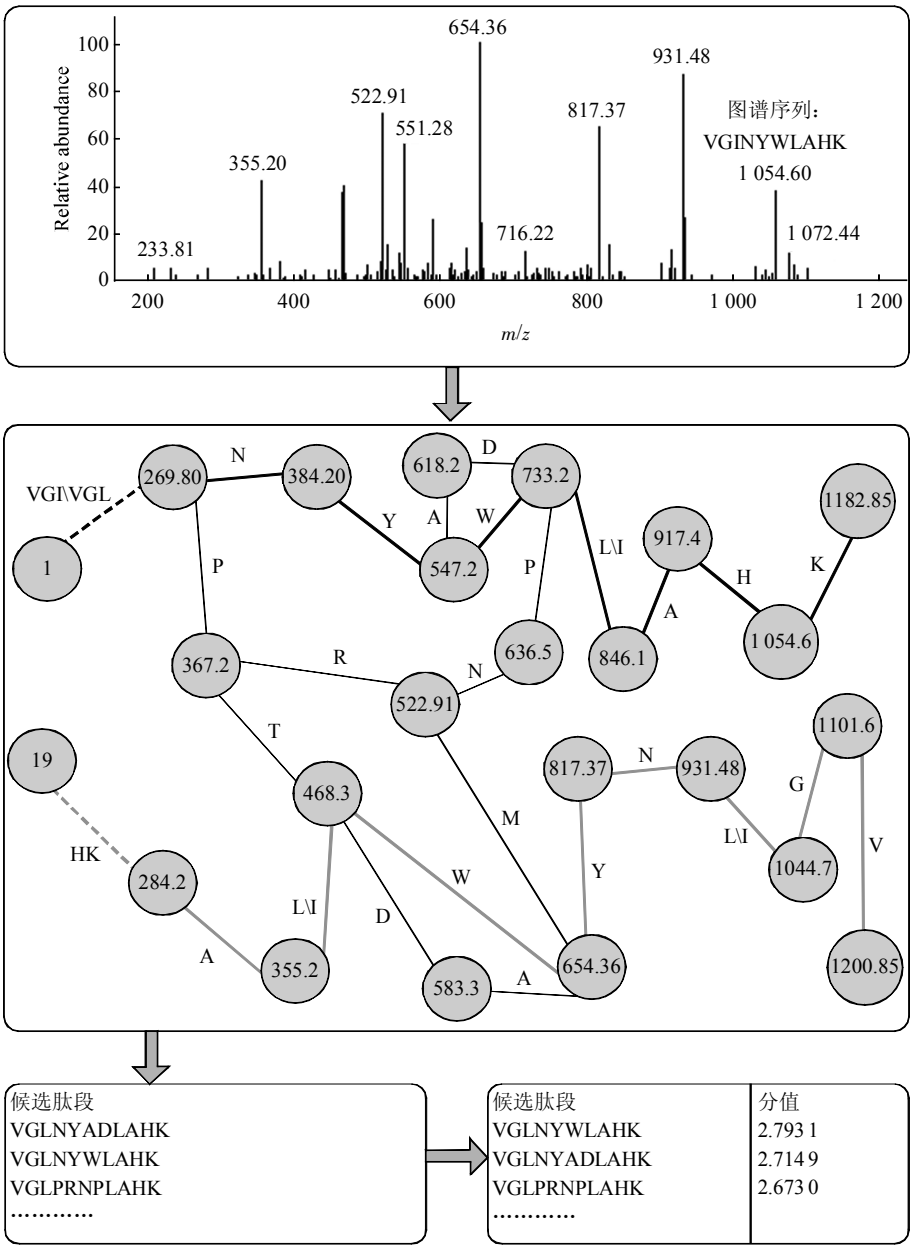

Fig. 4　The main workflow of graph theory based de novo peptide sequencing strategy

图 4　基于图论方法的基本流程图

在早期的基于图论的方法中，构建质谱峰连接图之前所采用的流程基本都是一致的，不同之处在于对质谱峰连接图的分析处理，例如，Fernandez-de-Cossio 等[23-24]提出的 SeqMS 方法，采用 Dijkstra 算法寻找从 N 端到 C 端的最短路径用以表示肽段序列，Taylor 等[25-27]提出的 Lutefisk 方法采用启发式的方法寻找 N 端到 C 端的最优路径，Dancik 等[28]提出的 Sherenga 方法采用一个有效的方法寻找质谱峰连接图中的一个反对称最长路径，Chen 等[29]提出的 Compute-Q 方法引入动态规划方法寻找质谱峰连接图中的最优路径，并且加快了图的搜索速度.

具体来讲，Lutefisk 方法首先将图谱数据简化，并确定图谱中显著离子的位置，然后寻找 N端和 C 端离子位置，之后再构建一个序列谱，x 轴代表质荷比，y 轴代表每个位点解离的可能性，接着从 N 端开始寻找 b 型离子系列的位点，得到候选序列之后通过一个打分函数进行排序，最后给出结果. Lutefisk 允许在得到的候选序列中存在空白区段，空白区段可能由若干个氨基酸残基组成. Sherenga 方法则能够从训练集中自动学习碎片离子类型和强度阈值，这种学习方法针对多种质谱仪均有效. Sherenga 构建的质谱峰连接图为一个有向无环图，使用一个似然比比值对经过规划得到的候选肽段序列打分. 随后，Havilio 等[30]又改进了 Sherenga 方法中的打分算法.

基于图论的从头测序方法一直在持续发展，到了 2005 年，Frank 等[31]提出了一个有效的从头测序方法 PepNovo. PepNovo 在 Havilio 等[30]的工作基础上作了进一步的改进，其核心是引入各种离子之间的关联可能性. PepNovo 使用两个似然比(假设检验)比值作为单个峰对应的分值，分子(第一个假设检验)是已知的肽段解离规则对应的似然比，分母(第二个假设检验)是随机解离过程对应的似然比，同时，PepNovo 引入概率网络模型减小计算空间，加快计算进程. 引入的概率网络模型考虑了各种离子类型之间的关联性概率，肽段解离位置之间的关系以及侧翼氨基酸对解离位置的影响. PepNovo 通过训练集来训练模型参数，类似于一个专家系统，训练集不同，得到的参数也不同. 同年，Grossmann 等[32]提出了一个包含启发式模块的从头测序工具 AUDENS，AUDENS 允许用户与软件交互，用户可以指定输入图谱峰的阈值和相关系数用以对图谱进行预处理.

2007 年，Mo 等[33]提出了一个新的从头测序方法 MSNovo. MSNovo 能够支持多种类型仪器产出的数据，能够支持+1、+2 和+3 价的母离子. MSNovo 引入了一个新的打分机制，同时结合质谱矩阵，使用动态规划方法解决从头测序问题. 实质上，质谱矩阵是质谱峰连接图的一个推广. 根据作者的报道，MSNovo 在多个数据集上都比以前的从头测序方法表现得更好. DiMaggio 等[34-35]将从头测序问题抽象为多约束条件的多目标优化问题，他们使用整数线性最优化来解决这个多目标优化问题，该方法对应的软件 PILOT 在两个小数据集上表现较好.

2010 年，Chi 等[36]提出了一个用于高能碰撞解离(higher-energy collisional dissociation, HCD)类型数据的从头测序算法 pNovo. 根据作者报道，使用数据库搜索方法鉴定出的 HCD 图谱中，80%以上都能够被 pNovo 正确测序.

另外，基于图论的方法还有 EigenMS[37]、PRIME[38-39]、Sub-denovo[40]、Liu 等[41]提出的树分解方法以及 Ning 等[42]对反对称模型的改进等.

基于图论方法的一个很大优点是能够将搜索空间简化为线性空间，使得从头测序方法能够在大规模的数据上得到应用. 这种方法的一个固有缺陷是不能有效分辨图谱中信号模糊区域的 b 型离子和 y 型离子，可能会导致测序存在错误.

2.3 综合方法

在解决串联质谱图谱从头测序问题的方法中，绝大多数是基于图论的方法，但研究者们还是提出了一些其他类型的方法. Ma 等[43-44]提出的 PEAKS 方法是这一类方法中的代表. PEAKS 一共包含四步：第一步是图谱预处理，即图谱噪声过滤和图谱峰聚合. 第二步是在简化后的图谱中使用类似于穷举的方法产生可能的候选肽段. PEAKS 首先通过每个峰周围的峰来计算该峰对应的 y 离子匹配分值和 b 离子匹配分值，如果该峰附近没有其他的峰，就赋予该峰一个惩罚分值，然后通过计算该峰附近的氨基酸残基使得该峰对应的总分值(b 离子匹配分值和 y 离子匹配分值)最大，并只保留 10 000 个最好的候选肽段，保留的候选肽段使用一个基于经验的并且更加精确的打分方法来评估，同时加入亚胺离子、中性丢失离子和内部解离离子的分值. 第三步和第四步为综合打分的差异分析以及分值正则化.

与 PEAKS 类似的一个方法是 RAId[45]. RAId

主要包含两个步骤：第一步是通过类似于 PEAKS 的策略得到一组候选肽段列表，然后通过候选肽段与实验图谱中解离位置和强度的匹配分值对候选肽段分级分类．第二步是将排名靠前的候选肽段与从蛋白质数据库中提前计算出的肽段列表进行匹配，如果有候选肽段匹配到肽段列表中的某个肽段就会生成报告，如果所有排名靠前的候选肽段都没有匹配到显著相关的肽段，那么就认为该图谱对应的肽段可能是数据库中不存在的新肽段．

Fischer 等[46]将隐马尔科夫模型(hidden markov model，HMM)应用到从头测序问题中，该方法对应的软件是 NovoHMM．在 NovoHMM 中，肽段序列被认为是隐含状态，而图谱则被认为是观测到的状态，改进的隐马尔科夫模型被用来从图谱直接推导肽段序列．NovoHMM 的一个优点是其打分规则由转移概率组成，即某个氨基酸残基跟随另一个氨基酸残基出现的概率组成的矩阵．需要强调的是，如果训练集过小，会导致 NovoHMM 在测试集上表现不好．

Spengler[47]提出了一个针对高精度仪器产出数据的综合方法，该方法包含两步．第一步是确定一组精确的碎片离子位置和目标肽段对应的氨基酸组合．第二步是对每个候选肽段计算期望的碎片离子信号，然后同观测到的碎片离子匹配．只有肽段对应的候选肽段数量比较少时，才能保证该方法的计算速度．该方法只适用于具有超高精度的仪器产出的数据．

针对低精度仪器产出的数据，Zhang[48]提出了一个名为 DACSIM 的方法．该方法包含两步，第一步采用一个分治策略[49-50]从图谱中产生一组候选肽段，第二步使用这些候选肽段产生理论图谱，然后将理论图谱与实验图谱做比对．该方法的一个不足是理论图谱的生成是基于经验的，并且是与仪器相关的．

另外还有一些方法可以归为综合方法，例如，Bruni 等[51]使用的组合数学模型，Zhang 等[52]提出的二维碎片关系模型，Demine 等[53]提出的适用于 MALDI 类型数据的 Sequit 方法，Heredia-Langner 等[54]提出的基于遗传算法的方法，Kanazawa 等[55]提出的使用离子峰强度和氨基酸解离位点强度比来进行从头测序的方法．

多数情况下，这一类方法是与仪器相关的，并且常常与肽段序列标签方法联用．

3 常用算法评估指标及数据集

研究者们已经提出了很多串联质谱图谱从头测序方法，比较各种方法的优劣也就成了从头测序问题研究中的一个子问题．Pevtsov 等[56]和 Pitzer 等[57]讨论了算法评估标准，分别对一些从头测序方法进行了比较．还有一些研究工作[31, 33-34, 46]也提出了一些算法评估指标．本节将介绍这些用于从头测序方法评估的指标和数据集．

第一个指标为完全正确率，也称为精确度．假定数据集中含有 m 张图谱，某种方法计算正确的图谱为 n 张，则定义完全正确率 I_1 为：

$$I_1=\frac{n}{m} \tag{1}$$

第二个指标为共同子序列正确率．假定数据集中含有 m 张图谱，共同子序列的长度为 k，给定 k 的一个范围，例如，$3\leqslant k\leqslant 10$，当 k 为某个固定值时，初始化 $t_k=0$，比对通过算法计算出的肽段和相应的真实肽段，如果计算出的肽段包含有真实肽段的长度为 k 的子序列，则 $t_k=t_k+1$，对 m 张图谱都重复上述过程，则定义共同子序列正确率 I_2 为：

$$[I_2]_k=\frac{t_k}{m} \tag{2}$$

第三个指标为氨基酸位点正确率．假定数据集中含有 m 张图谱，每张图谱对应的真实肽段为 $P_i(1\leqslant i\leqslant m)$，$P_i$ 包含的氨基酸残基的个数为 L_{p_i}，通过计算得到的肽段为 $CP_i(1\leqslant i\leqslant m)$，$CP_i$ 包含的氨基酸残基的个数为 L_{cp_i}，CP_i 包含正确的氨基酸残基的个数为 LC_{cp_i}，则定义氨基酸位点正确率 I_3 和 I_3' 为：

$$I_3=\frac{\sum_{i=1}^{m} LC_{cp_i}}{\sum_{i=1}^{m} L_{cp_i}} \tag{3}$$

$$I_3'=\frac{\sum_{i=1}^{m} LC_{cp_i}}{\sum_{i=1}^{m} L_{p_i}} \tag{4}$$

表 1 介绍了常用的用于评估和比较各种算法的数据集，包括数据集简称、图谱数目、图谱价态、作者联系方式和支持文献．

Table 1 Frequently used datasets for assessment of different *de novo* peptide sequencing algorithms

表 1 常用的串联质谱图谱从头测序算法评估数据集

数据集简称	图谱数目	图谱价态	作者联系方式	参考文献
ISB769	769	+2	arf@cs.ucsd.edu	[31, 33]
OPD280	280	+2	arf@cs.ucsd.edu	[31, 33, 46]
HUPO513	513	+2	tingchen@usc.edu	[33]
LTQ600	600	+2	tingchen@usc.edu	[33]
ISB646	646	+3	tingchen@usc.edu	[33]
IonTrap(LCQ)	36	+2	floudas@titan.princeton.edu	[34]
QTOF	38	+2	floudas@titan.princeton.edu	[34]
LTQ-FT	376	+2	ppevzner@cs.ucsd.edu	[58]

4 串联质谱图谱从头测序方法的应用

在实际应用中，由于存在样品污染、同位素峰干扰、肽段不完全解离导致的多数图谱中出现的重要 *b* 离子峰或者 *y* 离子峰丢失、N 端和 C 端信息缺失以及各种各样的噪声干扰等种种原因，从头测序方法的精度还不够高，难以得到正确的肽段序列信息，因此，单独应用从头测序方法还比较少见. 从头测序方法的主要应用策略是与同源性搜索、数据库搜索方法或者肽段序列标签方法组合起来使用.

例如，Waridel 等[59]使用 PepNovo[31]、Mascot[60]、序列相似性搜索对未测序的物种进行蛋白质序列分析，有效地鉴定出了新蛋白质的序列. Bandeira 等[61]使用一种组合方法对单克隆抗体进行测序，对比埃德曼降解法，显著加快了蛋白质测序速度. Ng 等[62]使用多级质谱分析策略和图谱比对算法对环状的非核糖体肽段进行测序，并且有效区分出了新分离出的非核糖体肽和已鉴定的非核糖体肽.

一些商业化的从头测序软件也得到了应用，例如，Kim 等[63]在研究一种基因组未测序的土壤细菌时，从 Mascot 搜库没有匹配到结果的图谱中人工挑选出质量较好的图谱，使用 PepSeq(Micromass, Manchester，UK)得到部分肽序列，通过与 Mascot 结果比较，再经 BLAST 同源性搜索，得到了高可信度的蛋白质以及一些新蛋白质. Lilla 等[64]也使用从头测序软件 PepSeq 和同源性比较程序 Blast-p 鉴定了一个新蛋白质.

5 总结与展望

本文介绍了串联质谱图谱从头测序问题，总结现有的从头测序算法，分析每类算法的优缺点，介绍了常用的算法评估指标和数据集，并简要介绍从头测序方法的一些应用，总结了部分软件的特性(表 2)和从头测序软件的网络资源(表 3).

Table 2 Characteristics of some *de novo* peptide sequencing tools

表 2 串联质谱图谱从头测序软件特性介绍

软件	在线 [1)]	类型	数据支持类型	PTM 支持 [2)]	母离子电荷支持 [3)]
Lutefisk	Y	命令行	IonTrap, QTOF	Y	+2
RAId	Y	命令行	IonTrap	Y	
PepNovo	Y	命令行	IonTrap, QTOF, FTMS	Y	+1, +2, +3
NovoHMM	Y	用户界面	LCQ	N	+2
PEAKS	Y	用户界面	IonTrap, QTOF, FTMS, MALDI	Y	+1, +2, +3
SeqMS	Y	用户界面	IonTrap	N	+1, +2
MSNovo	–	–	IonTrap	–	+2, +3
Sub-denovo	Y	命令行	IonTrap	N	+2
PILOT	–	–	IonTrap, QTOF	–	+2
AUDENS	Y	用户界面	IonTrap	N	+2

[1)] 在线是指提供网络下载链接或者提供 web server 服务, Y(支持), –(未知), N(不支持); [2)] 软件是否支持鉴定带有修饰的肽段, Y(支持), –(未知)，N(不支持), PTM(Post Translational Modification，翻译后修饰); [3)] 母离子常见的电荷数为+1, +2 和+3.

Table 3 Internet resources and references of *de novo* peptide sequencing tools

表 3 串联质谱图谱从头测序软件网络资源及参考文献

软件	网址	参考文献	免费
Lutefisk	http://www.hairyfatguy.com/Lutefisk/	[25−27]	Y
RAId	ftp://ftp.ncbi.nih.gov/pub/yyu/Proteomics/MSMS/RAId/Package/	[45]	Y
Compute-Q	−	[29]	−
PepNovo	http://cseweb.ucsd.edu/groups/bioinformatics/software.html#pepnovo	[31]	Y
NovoHMM	http://people.inf.ethz.ch/befische/proteomics/	[46]	Y
PILOT	−	[34]	−
SeqMS	http://www.protein.osaka-u.ac.jp/rcsfp/profiling/Seqms/SeqMS.html	[23−24]	Y
PRIME	http://csbl.bmb.uga.edu/downloads/prime/prime.html	[38−39]	−
Sub-denovo	http://msms.cmb.usc.edu/sub/	[40]	Y
DACSIM		[48]	−
MSNovo	http://msms.usc.edu/supplementary/msnovo	[33]	−
AUDENS	http://www.ti.inf.ethz.ch/pw/software/audens/	[32]	Y
EigenMS	−	[37]	−
Sherenga	http://www.agilent.com	[28]	N
MassSeq	http://www.micromass.co.uk, Micromass		N
PepSeq	http://www.micromass.co.uk, Micromass		N
RapiDeNovo™	http://www.bdal.de/us/products/software/biotools/overview.html		N
BioAnalyst	http://www.absciex.com/mk/get/software_bioanalyst, AB SCIEX		N
DeNovoX	http://www.thermo.com, Thermofinnigan		N
Sequit	http://www.sequit.org/	[53]	N
SeqLab™	http://www.ssi.shimadzu.com/		N
PEAKS	http://www.bioinformaticssolutions.com, Bioinformatics Solutions Inc.	[43−44]	N

串联质谱图谱从头测序问题是计算蛋白质组学中一个较难解决的问题，从目前的发展来看，算法正确率还不够高，针对质量较好的图谱也只能达到30%至45%，这限制了从头测序方法的应用．本文认为，应该从以下几个方面更深入地研究和应用从头测序方法：

a．基于图论的方法依旧是从头测序方法中的主流方法，问题的关键是寻找更好的打分函数，对碎片离子峰进行有效区分．

b．充分利用高精度仪器产出的数据．高精度数据的误差很小，通过母离子误差校正能够达到5 ppm 以下，穷举法能够在高精度仪器产出的数据中发挥更多的作用，甚至能够有效区分谷氨酸和赖氨酸．

c．由于从头测序方法全序列测序成功率比较低，应该允许通过计算得到的序列中有空白区段．这个观点在 Kim 等[65−66]的研究中已经有所体现，例如，D[186]AAENTDAQK，其中 186 就表示一个氨基酸残基或者多个氨基酸残基的组合，得到的含有空白区段的肽段不会影响后续的分析和应用．此外，目前还没有统一的标准对空白区段进行约束，例如，空白区段的最高质量限制、占整个肽段质量的百分比、最大氨基酸个数等，这些都还有待规范．

d．改进图谱预处理．图谱预处理对串联质谱从头测序很重要，特别是图谱中的同位素峰和低丰度峰对从头测序影响很大．

e．重视利用新实验技术产出的数据．例如，碰撞诱导解离与电子捕获解离(electron capture dissociation, ECD)、电子转移解离(electron transfer dissociation, ETD)产生的图谱信息具有互补性，一些文献[67−68]已有所报道，但是研究还不够深入．利用特定的化学修饰加强某些类型离子的信号，使用稳定同位素标记方法加强某些重点碎片离子的信号量，都有可能会提高从头测序的正确率．

f．多级质谱策略的应用．使用多级质谱[69]对

串联质谱中的某些重点碎片离子进行分析，具备提高从头测序正确率的潜力.

g. 加强对+3 价图谱的支持. 目前，仅有少量从头测序软件支持+3 价图谱，而且正确率普遍较低，提高+3 价图谱的测序精度有利于增加从头测序方法的覆盖度和应用范围.

h. 加强从头测序方法计算结果的确认研究. 目前大部分从头测序方法仅仅针对已知肽段信息的数据集验证算法的精度，对算法计算结果的过滤规则和确认研究较少，加强这个方面的研究有利于加快从头测序方法的应用步伐.

参 考 文 献

[1] James P. Protein identification in the post-genome era: the rapid rise of proteomics. Q Rev Biophys, 1997, **30**(4): 279-331

[2] Anderson N L, Anderson N G. Proteome and proteomics: new technologies, new concepts, and new words. Electrophoresis, 1998, **19**(11): 1853-1861

[3] Blackstock W P, Weir M P. Proteomics: quantitative and physical mapping of cellular proteins. Trends Biotechnol, 1999, **17**(3): 121-127

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

[5] Binz P A, Hochstrasser D F, Appel R D. Mass spectrometry-based proteomics: current status and potential use in clinical chemistry. Clin Chem Lab Med, 2003, **41**(12): 1540-1551

[6] Pusch W, Flocco M T, Leung S M, *et al*. Mass spectrometry-based clinical proteomics. Pharmacogenomics, 2003, **4**(4): 463-476

[7] Kolialexi A, Mavrou A, Spyrou G, *et al*. Mass spectrometry-based proteomics in reproductive medicine. Mass Spectrom Rev, 2008, **27**(6): 624-634

[8] Frobel J, Lehr S, Haas R, *et al*. Mass spectrometry-based proteomics and its potential use in haematological research. Arch Physiol Biochem, 2009, **115**(5): 286-297

[9] Domon B, Aebersold R. Mass spectrometry and protein analysis. Science, 2006, **312**(5771): 212-217

[10] Nesvizhskii A I, Vitek O, Aebersold R. Analysis and validation of proteomic data generated by tandem mass spectrometry. Nat Methods, 2007, **4**(10): 787-797

[11] Marcotte E M. How do shotgun proteomics algorithms identify proteins?. Nat Biotech, 2007, **25**(7): 755-757

[12] Steen H, Mann M. The ABC's (and XYZ's) of peptide sequencing. Nat Rev Mol Cell Biol, 2004, **5**(9): 699-711

[13] Sakurai T, Matsuo T, Matsuda H, *et al*. PAAS 3: A computer program to determine probable sequence of peptides from mass spectrometric data. Biological Mass Spectrometry, 1984, **11** (8): 396-399

[14] Hamm C W, Wilson W E, Harvan D J. Peptide sequencing program. Comput Appl Biosci, 1986, **2**(2): 115-118

[15] Ishikawa K, Niwa Y. Computer-aided peptide sequencing by fast atom bombardment mass spectrometry. Biological Mass Spectrometry, 1986, **13**(7): 373-380

[16] Siegel M M, Bauman N. An efficient algorithm for sequencing peptides using fast atom bombardment mass spectral data. Biological Mass Spectrometry, 1988, **15**(6): 333-343

[17] Johnson R S, Biemann K. Computer program (SEQPEP) to aid in the interpretation of high-energy collision tandem mass spectra of peptides. Biomed Environ Mass Spectrom, 1989, **18**(11): 945-957

[18] Zidarov D, Thibault P, Evans M J, *et al*. Determination of the primary structure of peptides using fast atom bombardment mass spectrometry. Biological Mass Spectrometry, 1990, **19**(1): 13-26

[19] Olson M T, Epstein J A, Yergey A L. *De novo* peptide sequencing using exhaustive enumeration of peptide composition. J Am Soc Mass Spectrom, 2006, **17**(8): 1041-1049

[20] Lu B, Chen T. Algorithms for *de novo* peptide sequencing using tandem mass spectrometry. Drug Discovery Today: BIOSILICO, 2004, **2**(2): 85-90

[21] Xu C, Ma B. Software for computational peptide identification from MS-MS data. Drug Discov Today, 2006, **11**(13-14): 595-600

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

[23] Fernandez-de-Cossio J, Gonzalez J, Besada V. A computer program to aid the sequencing of peptides in collision-activated decomposition experiments. Comput Appl Biosci, 1995, **11**(4): 427-434

[24] Fernandez-de-Cossio J, Gonzalez J, Betancourt L, *et al*. Automated interpretation of high-energy collision-induced dissociation spectra of singly protonated peptides by 'SeqMS', a software aid for *de novo* sequencing by tandem mass spectrometry. Rapid Commun Mass Spectrom, 1998, **12**(23): 1867-1878

[25] Taylor J A, Johnson R S. Sequence database searches *via de novo* peptide sequencing by tandem mass spectrometry. Rapid Commun Mass Spectrom, 1997, **11**(9): 1067-1075

[26] Taylor J A, Johnson R S. Implementation and uses of automated *de novo* peptide sequencing by tandem mass spectrometry. Anal Chem, 2001, **73**(11): 2594-2604

[27] Johnson R S, Taylor J A. Searching sequence databases *via de novo* peptide sequencing by tandem mass spectrometry. Mol Biotechnol, 2002, **22**(3): 301-315

[28] Dancik V, Addona T A, Clauser K R, *et al*. *De novo* peptide sequencing *via* tandem mass spectrometry. J Comput Biol, 1999, **6**(3-4): 327-342

[29] Chen T, Kao M Y, Tepel M, *et al*. A dynamic programming approach to *de novo* peptide sequencing *via* tandem mass spectrometry. J Comput Biol, 2001, **8**(3): 325-337

[30] Havilio M, Haddad Y, Smilansky Z. Intensity-based statistical scorer for tandem mass spectrometry. Anal Chem, 2003, **75** (3): 435-444

[31] Frank A, Pevzner P. PepNovo: *de novo* peptide sequencing *via* probabilistic network modeling. Anal Chem, 2005, **77**(4): 964-973

[32] Grossmann J, Roos F F, Cieliebak M, *et al*. AUDENS: a tool for automated peptide *de novo* sequencing. J Proteome Res, 2005, **4**(5):

1768−1774

[33] Mo L, Dutta D, Wan Y, *et al*. MSNovo: a dynamic programming algorithm for *de novo* peptide sequencing *via* tandem mass spectrometry. Anal Chem, 2007, **79**(13): 4870−4878

[34] DiMaggio P A, Jr, Floudas C A. *De novo* peptide identification *via* tandem mass spectrometry and integer linear optimization. Anal Chem, 2007, **79**(4): 1433−1446

[35] DiMaggio P A, Jr, Floudas C A, Lu B, *et al*. A hybrid method for peptide identification using integer linear optimization, local database search, and quadrupole time-of-flight or OrbiTrap tandem mass spectrometry. J Proteome Res, 2008, **7**(4): 1584−1593

[36] Chi H, Sun R X, Yang B, *et al*. pNovo: *de novo* peptide sequencing and identification using HCD spectra. J Proteome Res, 2010, **9**(5): 2713−2724

[37] Bern M, Goldberg D. *De novo* analysis of peptide tandem mass spectra by spectral graph partitioning. J Comput Biol, 2006, **13**(2): 364−378

[38] Yan B, Qu Y, Mao F, *et al*. PRIME: A mass spectrum data mining tool for *de novo* sequencing and PTMs identification. J Comp Sci Tech, 2005, **20**(4): 483−490

[39] Yan B, Pan C, Olman V N, *et al*. A graph-theoretic approach for the separation of b and y ions in tandem mass spectra. Bioinformatics, 2005, **21**(5): 563−574

[40] Lu B, Chen T. A suboptimal algorithm for *de novo* peptide sequencing *via* tandem mass spectrometry. J Comput Biol, 2003, **10**(1): 1−12

[41] Liu C, Song Y, Yan B, *et al*. Fast *de novo* peptide sequencing and spectral alignment *via* tree decomposition//Pac Symp Biocomput, Hawaii, 2006: 255−266

[42] Ning K, Leong H W. Algorithm for peptide sequencing by tandem mass spectrometry based on better preprocessing and anti-symmetric computational model. Comput Syst Bioinformatics Conf, 2007, **6**: 19−30

[43] Ma B, Zhang K, Hendrie C, *et al*. PEAKS: powerful software for peptide *de novo* sequencing by tandem mass spectrometry. Rapid Commun Mass Spectrom, 2003, **17**(20): 2337−2342

[44] Ma B, Zhang K, Liang C. An effective algorithm for peptide *de novo* sequencing from MS/MS spectra. J Comp Sys Sci, 2005, **70**(3): 418−430

[45] Alves G, Yu Y K. Robust accurate identification of peptides (RAId): deciphering MS2 data using a structured library search with *de novo* based statistics. Bioinformatics, 2005, **21**(19): 3726−3732

[46] Fischer B, Roth V, Roos F, *et al*. NovoHMM: a hidden Markov model for *de novo* peptide sequencing. Anal Chem, 2005, **77**(22): 7265−7273

[47] Spengler B. *De novo* sequencing, peptide composition analysis, and composition-based sequencing: a new strategy employing accurate mass determination by fourier transform ion cyclotron resonance mass spectrometry. J Am Soc Mass Spectrom, 2004, **15**(5): 703−714

[48] Zhang Z. *De novo* peptide sequencing based on a divide-and-conquer algorithm and peptide tandem spectrum simulation. Anal Chem, 2004, **76**(21): 6374−6383

[49] Zhang Z. Prediction of low-energy collision-induced dissociation spectra of peptides. Anal Chem, 2004, **76**(14): 3908−3922

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

[51] Bruni R, Gianfranceschi G, Koch G. On peptide *de novo* sequencing: a new approach. J Pept Sci, 2005, **11**(4): 225−234

[52] Zhang Z, McElvain J S. *De novo* peptide sequencing by two-dimensional fragment correlation mass spectrometry. Anal Chem, 2000, **72**(11): 2337−2350

[53] Demine R, Walden P. Sequit: software for *de novo* peptide sequencing by matrix-assisted laser desorption/ionization post-source decay mass spectrometry. Rapid Commun Mass Spectrom, 2004, **18**(8): 907−913

[54] Heredia-Langner A, Cannon W R, Jarman K D, *et al*. Sequence optimization as an alternative to *de novo* analysis of tandem mass spectrometry data. Bioinformatics, 2004, **20**(14): 2296−2304

[55] Kanazawa M, Anyoji H, Ogiwara A, *et al*. *De novo* peptide sequencing using ion peak intensity and amino acid cleavage intensity ratio. Bioinformatics, 2007, **23**(9): 1068−1072

[56] Pevtsov S, Fedulova I, Mirzaei H, *et al*. Performance evaluation of existing *de novo* sequencing algorithms. J Proteome Res, 2006, **5**(11): 3018−3028

[57] Pitzer E, Masselot A, Colinge J. Assessing peptide *de novo* sequencing algorithms performance on large and diverse data sets. Proteomics, 2007, **7**(17): 3051−3054

[58] Frank A M, Savitski M M, Nielsen M L, *et al*. *De novo* peptide sequencing and identification with precision mass spectrometry. J Proteome Res, 2007, **6**(1): 114−123

[59] Waridel P, Frank A, Thomas H, *et al*. Sequence similarity-driven proteomics in organisms with unknown genomes by LC-MS/MS and automated *de novo* sequencing. Proteomics, 2007, **7**(14): 2318−2329

[60] Perkins D, Pappin D, Creasy D, *et al*. Probability-based protein identification by searching sequence database using mass spectrometry data. Electrophoresis, 1999, **20**(18): 3551−3567

[61] Bandeira N, Pham V, Pevzner P, *et al*. Automated *de novo* protein sequencing of monoclonal antibodies. Nat Biotechnol, 2008, **26**(12): 1336−1338

[62] Ng J, Bandeira N, Liu W T, *et al*. Dereplication and *de novo* sequencing of nonribosomal peptides. Nat Methods, 2009, **6**(8): 596−599

[63] 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

[64] 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 Spectrom, 2005, **40**(3): 405−412

[65] Kim S, Gupta N, Bandeira N, *et al*. Spectral dictionaries:

Integrating *de novo* peptide sequencing with database search of tandem mass spectra. Mol Cell Proteomics, 2009, **8**(1): 53-69

[66] Kim S, Bandeira N, Pevzner P A. Spectral profiles, a novel representation of tandem mass spectra and their applications for *de novo* peptide sequencing and identification. Mol Cell Proteomics, 2009, **8**(6): 1391-1400

[67] Bertsch A, Leinenbach A, Pervukhin A, *et al*. *De novo* peptide sequencing by tandem MS using complementary CID and electron transfer dissociation. Electrophoresis, 2009, **30**(21): 3736-3747

[68] Datta R, Bern M. Spectrum fusion: using multiple mass spectra for *de novo* Peptide sequencing. J Comput Biol, 2009, **16**(8): 1169-1182

[69] Bandeira N, Olsen J V, Mann J V, *et al*. Multi-spectra peptide sequencing and its applications to multistage mass spectrometry. Bioinformatics, 2008, **24**(13): i416-i423

Algorithm Development of *de novo* Peptide Sequencing *Via* Tandem Mass Spectrometry*

SUN Han-Chang[1)], ZHANG Ji-Yang[1)], LIU Hui[1)], ZHANG Wei[1)], XU Chang-Ming[1)], MA Hai-Bin[1)], ZHU Yun-Ping[2)], XIE Hong-Wei[1)**]

([1)] *Department of Automatic Control, College of Mechatronics and Automation, National University of Defense Technology, Changsha* 410073, China; [2)] *State Key Laboratory of Proteomics, Beijing Proteome Research Center, Beijing Institute of Radiation Medicine, Beijing* 102206, China)

Abstract High-throughput mass spectrometry-based proteomics is developing rapidly in recent years. A key and essential issue in proteomics data processing is to identify proteins *via* tandem mass spectra. *De novo* peptide sequencing approach is database independent, which is a distinct advantage compared to database searching approach, so it can be used to analyze the data of new organisms or unsequenced organisms. *De novo* peptide sequencing problem is briefly described at first, and then the state-of-the-art of this problem is introduced from different aspects, which include the strategies with their advantages and disadvantages, frequently used algorithms and tools, criteria for algorithm assessment, and frequently used datasets for algorithm comparison. At last, the characteristics of some algorithms are summarized and some possible improvements of *de novo* peptide sequencing algorithm design are proposed.

Key words peptide, tandem mass spectrometry, *de novo* sequencing, algorithm development, computational proteomics

DOI: 10.3724/SP.J.1206.2010.00226

*This work was supported by grants from National Basic Research Program of Research (2006CB910803, 2006CB910706, 2010CB912700), Hi-Tech Research and Development Program of China (2006AA02A312), National S&T Major Project (2008ZX10002-016, 2009ZX09301-002) and State Key Laboratory of Proteomics (SKLP-Y200811).

**Corresponding author.

Tel: 86-731-84576311, E-mail: xhwei65@nudt.edu.cn

Received: April 27, 2010 Accepted: July 5, 2010