PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2013, 40(1): 80~92
www.pibb.ac.cn

规模化蛋白质鉴定中母离子的准确检测技术研究 *

袁作飞[1, 2] 邬 龙[1, 2] 刘 超[1, 2] 迟 浩[1, 2] 樊盛博[1, 2]
张 昆[1, 2] 曾文锋[1, 2] 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)] 中国科学院大学，北京 100049)

摘要 蛋白质组学的基础研究之一是蛋白质鉴定．规模化的蛋白质鉴定通常采用“鸟枪法”，即选择一些酶切肽段(母离子)碎裂生成二级谱图，通过二级谱图及其母离子质量鉴定肽段，再推断对应的蛋白质．在鉴定过程中，母离子质量是一个关键参数．母离子是否是肽段的单同位素峰决定了正确肽段是否能进入候选，母离子的质量精度决定了候选肽段的数目．本文从判断单同位素峰和系统误差校准这两个角度研究了母离子的准确检测技术．判断单同位素峰的技术在蛋白质上已有研究，包括电荷判断、单同位素峰判断和重叠同位素峰判断．可以借鉴蛋白质水平的技术研究母离子的单同位素峰判断方法．同时母离子的系统误差校准也有较为成熟的方法．这两个角度的研究有助于提高规模化蛋白质的鉴定率．

关键词 蛋白质组学，蛋白质鉴定，母离子，单同位素峰，系统误差
学科分类号 Q51，TP39 **DOI**: 10.3724/SP.J.1206.2012.00167

蛋白质组学(proteomics)是研究细胞或组织内所有表达的蛋白质的一门基础学科．其研究内容包括蛋白质鉴定、翻译后修饰分析、定量分析、相互作用网络、结构预测、功能分类、疾病诊断与药物设计等，其中蛋白质鉴定是基础步骤[1]．研究蛋白质组需要规模化的方法．常用的规模化蛋白质鉴定方法是“鸟枪法”(shotgun)[2]．其基本思路是生物样品中的蛋白质被酶切成肽段，经过色谱分离进入质谱仪，质谱仪可以对某个时间点(洗脱时间或保留时间)的肽段离子进行扫描生成一级谱(横轴是离子的质荷比，纵轴是离子的强度)，在一级谱上以强度高的离子(母离子)为中心设置一个±1 Th 左右的窗口进行碎裂，扫描碎片离子生成二级谱．这就是传统的数据依赖(data dependent)的采集模式．

采集完二级谱之后，可以通过数据库搜索、谱库搜索、从头测序(de novo)等计算方法鉴定二级谱对应的肽段，再根据鉴定的肽段推断出蛋白质．鉴定肽段的过程包括通过母离子的质量及误差获得候选肽段、二级谱和候选肽段打分、候选肽段按分数排序，通常把第一名的肽段作为可能正确的肽段．肽段鉴定中，通常使用母离子的单同位素峰质量．如果质谱仪的数据处理软件没有导出正确的母离子电荷或单同位素峰质量，则正确肽段有可能进入不了候选而无法得到正确结果．如果能通过质谱仪或者计算方法获得非常小的母离子质量误差，则可以极大减少候选肽段的数目，从而提高肽段鉴定的速度．所以准确检测母离子，包括正确的单同位素峰和较小的质量误差，将提高规模化蛋白质鉴定的效率．

本文首先在第 1 部分对规模化蛋白质鉴定中母离子检测的两个问题做了描述，然后在第 2 部分分别介绍了蛋白质水平和肽段水平的单同位素峰判断方法，并对这两个水平下的技术特点做了比较，还介绍了在肽段水平做系统误差校准的方法；第 3 部分是对全文的总结和对未来的展望．

* 国家重点基础研究发展计划(973)(2010CB912701, 2012CB910602), 中国科学院知识创新计划(KGGX1-YW-13), 国家自然科学基金(30900262)和国家高技术研究发展计划(863)(2007AA02Z315, 2008AA02Z309)资助项目.
** 通讯联系人.
Tel: 010-62601016, E-mail: smhe@ict.ac.cn
收稿日期：2012-04-01，接受日期：2012-06-05

1　问题描述

在规模化的蛋白质鉴定中，质谱仪的发展速度很快．现在主流的质谱仪一级谱都是高精度的，比如 Orbitrap、FTICR、高精度 QTOF 等[3-5]．对于母离子即肽段离子，在高精度的一级谱上同位素峰之间能区分开，而低精度的一级谱则没法区分开同位素峰．在一个同位素峰簇中，质荷比最小的称为单同位素峰，第二小的称为第一同位素峰，以此类推．在高精度的一级谱上，如果在一个质荷比区域内没有干扰峰，直接根据同位素峰的质荷比间隔和同位素峰的强度分布可以获得母离子单同位素峰的质荷比和电荷．但是如果出现干扰峰，则情况就变得复杂．比如：在+2 电荷的两个峰中间出现一个峰，有可能判为+4 电荷(图 1a)；在单同位素峰前面的等间隔处出现一个峰，则会影响单同位素峰的质荷比判断(图 1b)；某两个同位素峰簇在某个或某些峰上重叠，则会影响单个同位素峰簇的强度分布，从而影响单同位素峰的判断(图 1c)．所以即使是高精度的一级谱，判断母离子的单同位素峰也不是一件简单的事情．

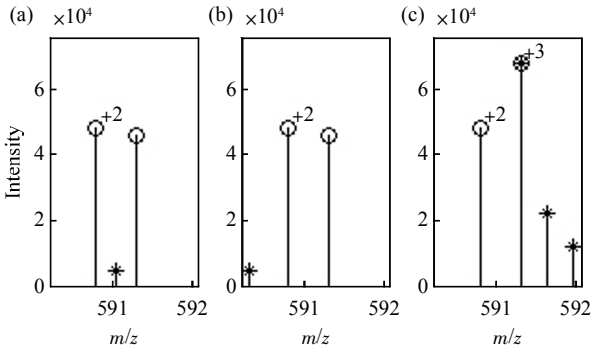

Fig. 1　Interference with isotopic clusters

图 1　同位素峰簇受干扰的示意图

(a) 电荷受干扰. (b) 单同位素峰受干扰. (c) 同位素峰簇重叠.

因为母离子的同位素峰簇容易受到干扰，准确判断单同位素峰比较困难，相应的算法比较复杂，计算时间长，准确率也不高．而对于质谱仪，要求其有较快的数据采集速度，所以不适合在数据采集的时候使用复杂的算法判断母离子的单同位素峰．传统的数据依赖的采集模式只采用了简单快速的方法，即在高强度峰的附近设置一个±1 Th 左右的窗口，把窗口内的峰都碎裂，扫描后生成二级谱．这样做有三点好处：速度快、单同位素峰也被碎裂、保证了二级谱的信噪比高．因为即使可以定位到单同位素峰，在其附近设置一个 ppm 量级的窗口，会导致二级谱的信噪比很低而无法被鉴定．

在输出母离子的质荷比和电荷时，质谱仪的数据处理软件如果没有判断单同位素峰，则默认输出的是碎裂窗口中心的峰．这个峰的特点是强度高，但不一定是单同位素峰．在实际的观察中发现，当肽段的质量达到 1 800 Da 左右时，第一同位素峰开始变得最高，当肽段的质量达到 3 300 Da 左右时，第二同位素峰开始变得最高．所以碎裂中心的峰可能不是单同位素峰．如果根据碎裂中心峰的质量，用 ppm 量级的误差搜索数据库，有可能得不到正确的候选肽段；如果用 Da 量级的误差搜索数据库，虽然可能得到正确的候选肽段，但损失了高精度过滤候选肽的能力．另外，碎裂窗口会引入干扰峰或共洗脱肽段(如图 1 所示，在 c 图中+2 电荷和+3 电荷肽段是共洗脱肽段，假设+3 电荷肽段在碎裂中心，后文提到的共洗脱肽段特指除碎裂中心外的肽段，如 c 图中的+2 电荷肽段)．干扰峰会影响碎裂中心的单同位素峰判断．共洗脱肽段的单同位素峰也要判断，因为在二级谱中共洗脱肽段也会碎裂，判断了共洗脱肽段的单同位素峰，就可以鉴定共洗脱肽段(图 2)．如果共洗脱肽段和碎裂中心的肽段有重叠峰，则情况变得更复杂，会影响各自单同位素峰的判断．以前的方法认为一个二级谱只是由一个肽段碎裂而成．由于碎裂窗口引入了共洗脱肽段，实际上二级谱可能是混合谱，即由多个肽段碎裂而成．以前的方法对于一个二级谱只给碎裂中心的单同位素峰，实际上需要给出碎裂窗口所有可能的单同位素峰．总之，传统数据依赖的采集模式很好地完成了数据采集的工作，把准确、全面判断单同位素峰的工作留给了后处理．

准确判断母离子的单同位素峰可以提高规模化蛋白质的鉴定率．而母离子的系统误差校准则可以提高规模化蛋白质鉴定的速度和可信度．在规模化的蛋白质鉴定中，质谱仪在生成一级谱的过程中受环境温度、时间、肽段的质量和色谱峰的强度等因素的影响，最后扫描到的肽段质荷比存在系统误差[6]．在高精度的质谱仪中，系统误差的存在较为明显．如果校准的周期太长，比如一周，系统误差会偏离几个 ppm．因此，对数据质量要求高的生物实验室每天做一次校准，保证系统误差在 0 ppm 附

近．对于系统误差偏离较大的实验数据，则需要通过后处理对系统误差做校准．不管是通过质谱仪还是后处理做系统误差校准，校准后的数据对后续的规模化蛋白质鉴定都有帮助[7].

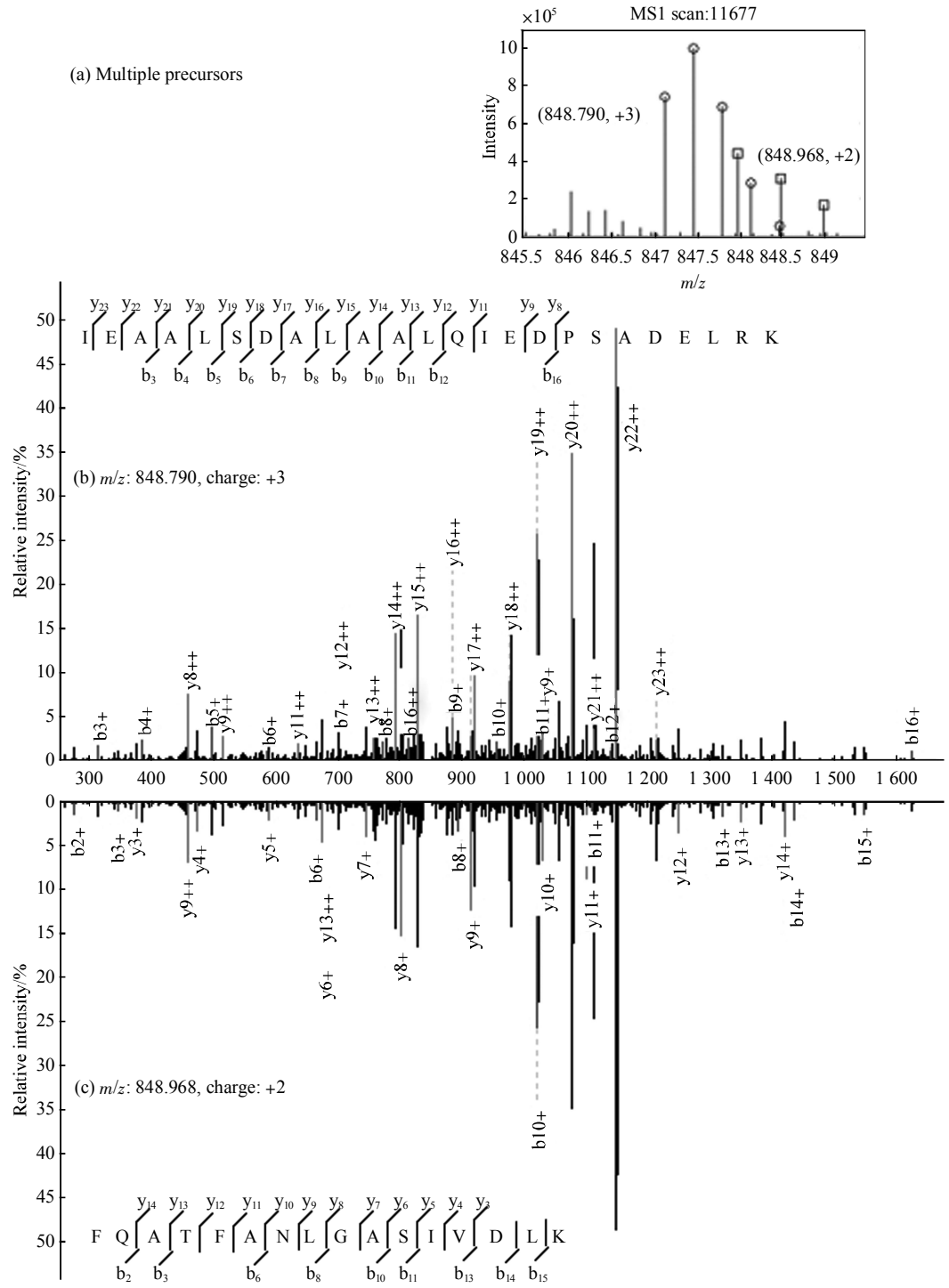

Fig. 2　Identification of co-eluted peptides

图 2　共洗脱肽段的鉴定

(a) 两个母离子的(质荷比, 电荷)对分别是(848.790, +3)和(848.968, +2). (b) 第一个母离子(848.790, +3)的鉴定肽段是 IEAALSDALAALQIEDP-SADELRK. (c) 第二个母离子(848.968, +2)的鉴定肽段是 FQATFANLGASIVDLK.

2 已有方法

由于首先在蛋白质水平判断了单同位素峰，其中的一些方法对肽段水平判断单同位素峰也产生了影响，所以下面将从蛋白质和肽段两个层次综述单同位素峰的判断，均包括电荷判断、单同位素峰判断和重叠同位素峰判断三个方面，并对这两个层次的方法做比较．然后从质谱仪和后处理的角度综述母离子质量的系统误差校准．

2.1 判断蛋白质的单同位素峰

2.1.1 Zubarev 等的方法．判断单同位素峰的方法起源于蛋白质大分子．Zubarev 等[8]观察到了单同位素峰质量和平均质量的线性关系．拟合之后给出了如下的关系式：对于蛋白质分子，Mmono = (1462/1463)Mav；对于 DNA 和 RNA，Mmono = (2091/2092)Mav．这里假设电荷判断已经解决、无重叠同位素峰．这个模型虽然精确度不够，但可以用于估计单同位素峰质量，作为其他方法的参考．

2.1.2 Senko 等的方法．

由于大蛋白质的分子质量很大，远大于肽段的分子质量，导致大蛋白质的同位素峰数目很多，强度分布呈现高斯分布[9]．在蛋白质的同位素峰簇中单同位素峰的信噪比可能很低，和噪音混在一起无法识别．这是蛋白质大分子判断单同位素峰的第一个难点．由于蛋白质分子质量大，而质谱仪的扫描范围有限，为了能检测到蛋白质，质谱仪在离子化的过程中让蛋白质带了很高的电荷，比如几十甚至上百．所以需要先判断电荷，以便得到蛋白质的分子质量．由于电荷高，而质谱仪的分辨率有限，仅根据相邻峰的质荷比间隔难以判断电荷，所以需要一些方法来获得同位素峰簇的电荷[10]．这是蛋白质大分子判断单同位素峰的第二个难点．早期的实验中少数几个标准蛋白质已做过纯化，所以基本没有重叠同位素峰簇．而后来的实验通量高、蛋白质多，即使做过纯化，也可能出现重叠同位素峰簇．这是蛋白质大分子判断单同位素峰的第三个难点．

在早期的方法中，不考虑第三个难点．第二个难点也可以解决，比如 Labowsky 等[11]的方法，就可以得到同位素峰簇的电荷．这样把同位素峰簇中每个峰的质荷比转换成质量，用峰的强度做加权平均得到蛋白质的平均质量．这个蛋白质已有了实验同位素峰的强度分布．如果能估计其理论同位素峰的强度分布，则可以判断后者估计得是否准确．如果估计准确，则可以从理论同位素峰簇的分子式计算出单同位素峰的质量．

如何估计理论同位素峰的强度分布？通过数据库统计得到一个平均氨基酸，比如在 protein information resource protein sequence database (PIR-PSD)数据库中统计每个氨基酸的比例，作为权重加到每个氨基酸的分子式上，得到平均氨基酸 $C_{4.9384}H_{7.7583}N_{1.3577}O_{1.4773}S_{0.0417}$．假设蛋白质的分子式是这个平均氨基酸分子式的倍数．已经获得了蛋白质的平均质量和平均氨基酸的平均质量为 111.1254 Da，两者相除即可得到这个倍数．把这个倍数乘到平均氨基酸的分子式上，即可获得蛋白质的一个准分子式．由于蛋白质分子式中的元素个数是整数，所以要对这个准分子式进行调整．把 C、N、O、S 的浮点个数取整，小数部分的质量和加到 H 的个数上并取整．这样借助平均氨基酸，通过蛋白质的平均质量估计了一个分子式．有了这个分子式就可以求理论同位素峰的强度分布，比如 Yergey 等[12]的方法．

如何检查估计的理论同位素峰强度分布是准确的？首先可以求实验同位素峰的强度分布和理论同位素峰的强度分布的距离．只有一个距离值看不出是否准确．于是可以增减蛋白质的平均质量，每个平均质量可以估计出对应的分子式，计算出对应的理论同位素峰的强度分布，再和实验同位素峰的强度分布求距离．有了多个距离，就可以根据距离的大小判断是否准确．取最小距离对应的分子式，其理论同位素峰的强度分布被认为是最准确的．根据这个分子式就可以计算出单同位素峰的质量．这个方法就是 Senko 等[9]提出的经典平均氨基酸模型．

这个方法在 16941.97 Da 的马肌红蛋白上做测试，估计的单同位素峰质量为 16941.77 Da，比理论值少 0.2 Da．这个结果从一定程度上说明这个方法可靠．平均氨基酸模型的方法给我们提供了一个思路，可以借助平均氨基酸估计理论分子式和理论同位素峰强度分布，这样在不知道分子式的情况下也有一个同位素峰强度分布的参考，进而能和实验同位素峰强度分布比较，从而判断单同位素峰．不过，平均氨基酸模型的方法有一个基本假设，就是在由氨基酸组成的数据库下任何一个分子式是平均氨基酸分子式的倍数．这个假设在一些特殊情况下会失效，比如修饰、标记、糖结构等．这时平均氨基酸模型需要修改，需要把非氨基酸的部分单独拿

出来，氨基酸部分的分子式仍满足这个基本假设．在这些特殊情况下，如果没有或者没法对平均氨基酸模型做修改，则会引入误差．在有修饰的情况下，模型修正前后用平均氨基酸模型计算的同位素峰强度分布，分别和根据修饰肽段计算的理论分布比较，这样可以看出哪个估计方法和真实的理论分布更接近．

平均氨基酸模型的方法还有第二个假设，就是实验同位素峰强度分布稳定且准确．准确是指实验同位素峰强度分布和理论同位素峰强度分布非常接近，稳定是指两个分布的误差集中在一个很小的范围内．这个假设是基于质谱仪有良好的强度测量性能．正是基于这个假设，才可以用实验同位素峰强度分布判断估计的理论同位素峰强度分布是否准确．如果质谱仪的强度测量性能不是很好，则这个假设不成立，即实验同位素峰强度分布本身不准，最后估计的理论同位素峰强度分布也不准．可以用有鉴定结果的质谱数据，在一级谱上计算实验同位素峰强度分布，用鉴定结果计算理论同位素峰强度分布，计算两个分布的距离，统计这个距离的规律，从而可以检测质谱仪的强度测量性能，检验这个假设．

利用平均氨基酸模型判断单同位素峰还引出了两个问题：电荷判断和同位素峰簇重叠的判断．判断蛋白质电荷的算法复杂，计算时间较长．先判断蛋白质电荷再判断单同位素峰，整个过程的计算时间更长．另外，这里的蛋白质大分子是经过纯化的，没有同位素峰簇重叠的问题．总之，Senko 等提出的平均氨基酸模型可以较准确地判断蛋白质大分子的单同位素峰，前提条件是样品不要太复杂(比如没有修饰、标记等复杂情况)、质谱仪的性能较好(特别是强度测量性能)、蛋白质经过纯化(不会有重叠同位素峰簇)．和 Zubarev 等的方法相比，平均氨基酸模型利用了同位素强度分布，比只利用平均质量的信息更丰富，所以精确度更高．

2.1.3 Horn 等的方法．

后来的实验通量高，不可避免地出现了重叠同位素峰簇．如何处理这个问题，成为各个方法的焦点．Horn 等[13]提出了一个重叠同位素峰簇的解决方案，把判断单同位素峰的工作推向了实用化，并开发了相应的软件 THRASH．THRASH 利用 Patterson 和 FT 结合的方法判断电荷[10]，用平均氨基酸模型的方法判断单同位素峰[9]．和纯化蛋白质的实验同位素峰簇相比，高通量的实验同位素峰簇更加复杂，原因是有重叠同位素峰簇和噪音峰的影响，导致目标蛋白的电荷或者单同位素峰判错．针对噪音的情况，THRASH 引入了信噪比的计算，根据信噪比判断噪音，从而减少噪音的影响．针对重叠的同位素峰簇，THRASH 在实验同位素峰簇的基础上减去其对应的理论强度，剩下的谱峰再做单同位素峰判断，采取的是贪心策略．噪音和重叠的处理是 THRASH 和 Senko 等工作的不同之处．

由于 THRASH 考虑了单同位素峰判断的三个方面、工作全面、方法有代表性，因此产生了较大影响．一系列方法都采取了和 THRASH 相同的思路，只是细节上略有变化，比如 Xtract、AID-MS、MasSPIKE、MassSpec [14-17]．其中，Xtract 基于 THRASH，后三个在判断电荷和单同位素峰的方法上和 THRASH 不同．YADA 没有采用消减的贪心方法判断重叠[18]．MS-Deconv 则采用动态规划判断最优路径后再分解成多个分布来判断重叠[19]．

2.2 判断肽段的单同位素峰

2.2.1 Hoopmann 等的方法．

由于 THRASH 的影响力很大，很容易想到直接用 THRASH 判断肽段的单同位素峰，比如 ICR-2LS(http://ncrr.pnl.gov/software/)和 Decon2LS[20]．也可以采用类似 THRASH 的思路判断肽段的单同位素峰，比如 Gras 等、Breen 等、Isoconv、ESI-Isoconv、PepList、Noy 等[21-26]．前三个因为是在 MALDI 数据上处理，所以不需要判断电荷．另外，msInspect、DTASuperCharge 采用了平均氨基酸模型，但没有考虑重叠同位素峰簇[27-28]．由于 THRASH 主要以 Senko 等的工作为基础，所以 THRASH 的计算时间较长．对于肽段水平的单同位素峰判断来说，THRASH 不仅在计算性能上需要提高，比如减少电荷判断的时间，在算法上也有改进的地方，比如减少平均氨基酸模型的误差．

由于 THRASH 不够完善，Hoopmann 等[29]对平均氨基酸模型在肽段单同位素峰判断的实用上做了改进，并开发了相应的软件 Hardklor．第一点改进是根据相邻峰的质荷比间隔判断肽段的电荷．第二点改进是根据样品中的修饰、标记等修正平均氨基酸模型．这两点改进正是 THRASH 的不足之处．同时 Hardklor 也考虑了重叠同位素峰簇，它是看多个同位素峰簇的强度叠加后哪种组合和实验同位素峰簇最接近，也是贪心策略，和 THRASH 相减

的思路不一样．与 THRASH 相比，Hardklor 的计算速度更快，平均氨基酸模型的误差更小，精确度更高．Hardklor 是目前基于平均氨基酸模型的算法和软件中考虑最为全面的．

在肽段水平判断单同位素峰，采用 Hardklor 的组合方式判断重叠的还有 MATCHING、GISTool、RAAMS、Roussis 等[30-33]．和 THRASH、Hardklor 的贪心策略不同，还有非贪心的方法判断重叠．Bielow 等[34]采用整数线性规划解混合模型．Samuelsson 等[35]采用二次规划解混合模型．Du 等和 NITPICK 软件采用 LASSO 求最少肽段的变量模型[36-37]．McIlwain 等和 BPDA 软件采用 Bayesian 模型判断重叠[38-39]．

基于平均氨基酸模型的方法判断肽段的单同位素峰，除了 THRASH 和 Hardklor 为代表的软件，还有一些软件．DeconMSn 采用 Patterson/FT 和相邻质荷比间隔结合的方法判断高精度的电荷，用 SVM 判断低精度的电荷，用 Senko 等的方法判断单同位素峰，没有考虑重叠同位素峰簇[40]．Mascot Distiller 采用了 Gras 等的思路判断单同位素峰，即和 THRASH 的思路一样，缺点是速度慢、对同位素标记的准确率低、实验分布很差时准确率低、对于已经中心化的数据效果不好(http://www.matrixscience.com/distiller.html)．

有了这些判断肽段单同位素峰的方法，可以结合一级谱的质荷比、保留时间，或者鉴定结果对单同位素峰做校正．Shin 等[41]利用 ICR-2LS 导出所有一级谱的单同位素峰，然后根据质荷比和保留时间聚类，二级谱母离子的质荷比、保留时间和聚类结果做匹配，从而对母离子的单同位素峰做校正 (PE-MMR)．Jung 等[42]集成了三个软件：DeconMSn、修正的 PE-MMR、做系统误差校准的 DtaRefinery，进一步提高了单同位素峰判断的准确性(iPE-MMR)．Scherl 等[43]则利用 Hardklor 判断单同位素峰，把碎裂窗口和保留时间内的峰簇强度加起来作为分数排序，输出前 3 个单同位素峰．

2.2.2 Park 等的方法．除了用平均氨基酸模型的方法判断肽段单同位素峰，Park 等[44]则是根据相邻同位素峰的比值关系来判断单同位素峰，比如相邻同位素峰的强度比值是肽段质量的函数等．通过数据库，统计这些比值的规律，得到最大值、最小值和平均值三条线，观察实验强度比值和这三条线的距离，从而对实验强度比值打分，考虑把不同峰作为单同位素峰的情况，甚至考虑单同位素峰不存在的情况，看哪种情况最接近统计规律，即打分最好，就输出对应的峰作为单同位素峰．这个方法的优点是考虑到同位素峰簇采集的强度不高，单同位素峰可能没有出现．其不足之处是没有考虑重叠的同位素峰簇．

2.2.3 Cox 等的方法．Park 等已经注意到肽段可能在强度不高时被采集了，这时由于质谱仪的动态范围有限且谱峰强度低，可能没有采集到单同位素峰，或者实验同位素峰簇的强度分布和理论同位素峰簇的强度分布相差较大，不利于单同位素峰的判断．这是上述所有只利用单张一级谱来判断单同位素峰的方法遇到的信号缺失或失真的问题．最近在一级谱上重构色谱峰做定量的软件非常多，比如 MaxQuant、OpenMS、VIPER、Superhirn、MapQuant 和 MZmine[45-50]．其中 Cox 等[45]观察到噪音没有同位素峰簇，同一肽段的同位素峰在色谱上有相同强度变化，而不同肽段的色谱峰强度变化不一样(图 3)．所以色谱峰是区分肽段和噪音以及不同肽段的重要依据．对于同一肽段，可以根据质荷比间隔和色谱峰强度变化判断同位素峰簇，再根据实验同位素峰强度分布和平均氨基酸得到的同位素峰强度分布的相似度判断单同位素峰. 这是 MaxQuant 判断单同位素峰的部分．可以看出 MaxQuant 利用了同位素峰强度分布的相似度和色谱峰的相似度，有效避免了低强度谱峰带来的信号缺失或失真的问题.

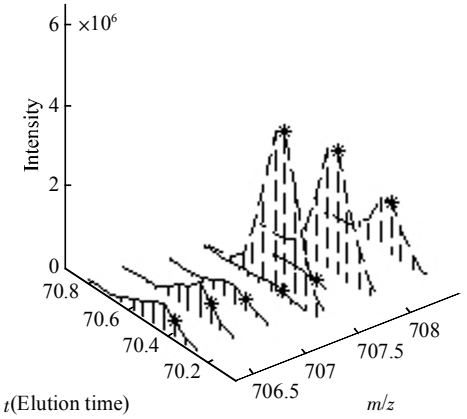

Fig. 3 Intensity of elution profiles

图 3 色谱峰强度变化的示例

2.2.4 质谱仪配套的数据处理工具．质谱仪厂商 Thermo Scientific 提供了和质谱仪配套的数据处理软件 Xcalibur(和质谱仪绑定)或者 MSFileReader，

它包含的控件提供两个接口：导出碎裂中心的峰[51]，对部分碎裂中心的峰做校准．前一个接口对应的软件有 RawXtract、MakeMS2、DTA Generator，后一个接口对应的软件有 Extract_MSn、ReAdW、Msconvert、pXtract，如表 1 所示．对质谱数据导出所有可能的单同位素峰，做数据库搜索鉴定、过滤得到可靠的标注集．观察标注集上 Extract_MSn 导出的母离子是否是单同位素峰．发现 Extract_MSn 在两类情况下导出的不是单同位素峰(图 4)：a．碎裂中心最高，尽管前面还有同位素峰，但仍把碎裂中心作为单同位素峰导出；b．碎裂中心前面还有一个同位素峰簇，把前面这个同位素峰簇的某个峰作为单同位素峰导出．同样，观察标注集上 MaxQuant 导出的母离子是否是单同位素峰．发现 MaxQuant 在四类情况下导出的不是单同位素峰(图 5)：a．多个同位素峰簇交叉，只给了其中一个单同位素峰；b．碎裂中心前面还有一个同位素峰簇，把前面这个同位素峰簇的某个峰作为单同位素峰导出；c．单同位素峰判断正确，但电荷错误；d．单同位素峰的位置正确，但质量精度偏大．可见，两个广泛使用的数据处理软件在判断肽段单同位素峰时会出错．

Table 1　Software tools corresponding to the interface of mass spectrometers

表 1　和质谱仪接口相关的软件

软件	网址
MSFileReader	http://sjsupport.thermofinnigan.com/public/detail.asp?id=703
RawXtract	http://fields.scripps.edu/downloads.php
MakeMS2	http://proteome.gs.washington.edu/software/makems2/
DTA Generator	https://search.apcf.edu.au/dbdownloads/dta-generator.zip
Extract_MSn	http://sjsupport.thermofinnigan.com/public/detail.asp?id=701
ReAdW	http://sourceforge.net/projects/sashimi/files/
Msconvert	http://proteowizard.sourceforge.net/
pXtract	http://pfind.ict.ac.cn/pXtract.htm

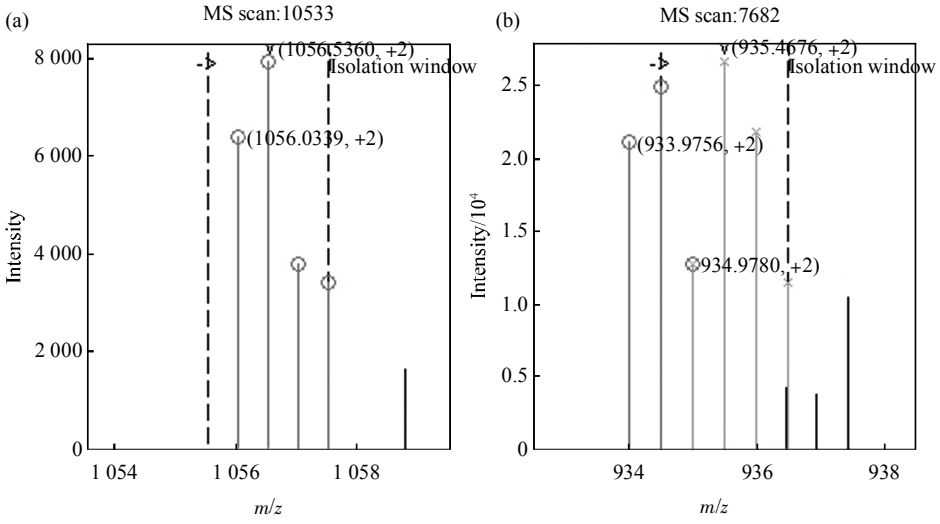

Fig. 4　Incorrect examples of Extract_MSn

图 4　Extract_MSn 判错的示例

(a) Extract_MSn 把最高峰作为单同位素峰导出. (b) Extract_MSn 只把最左边的峰作为单同位素峰导出.

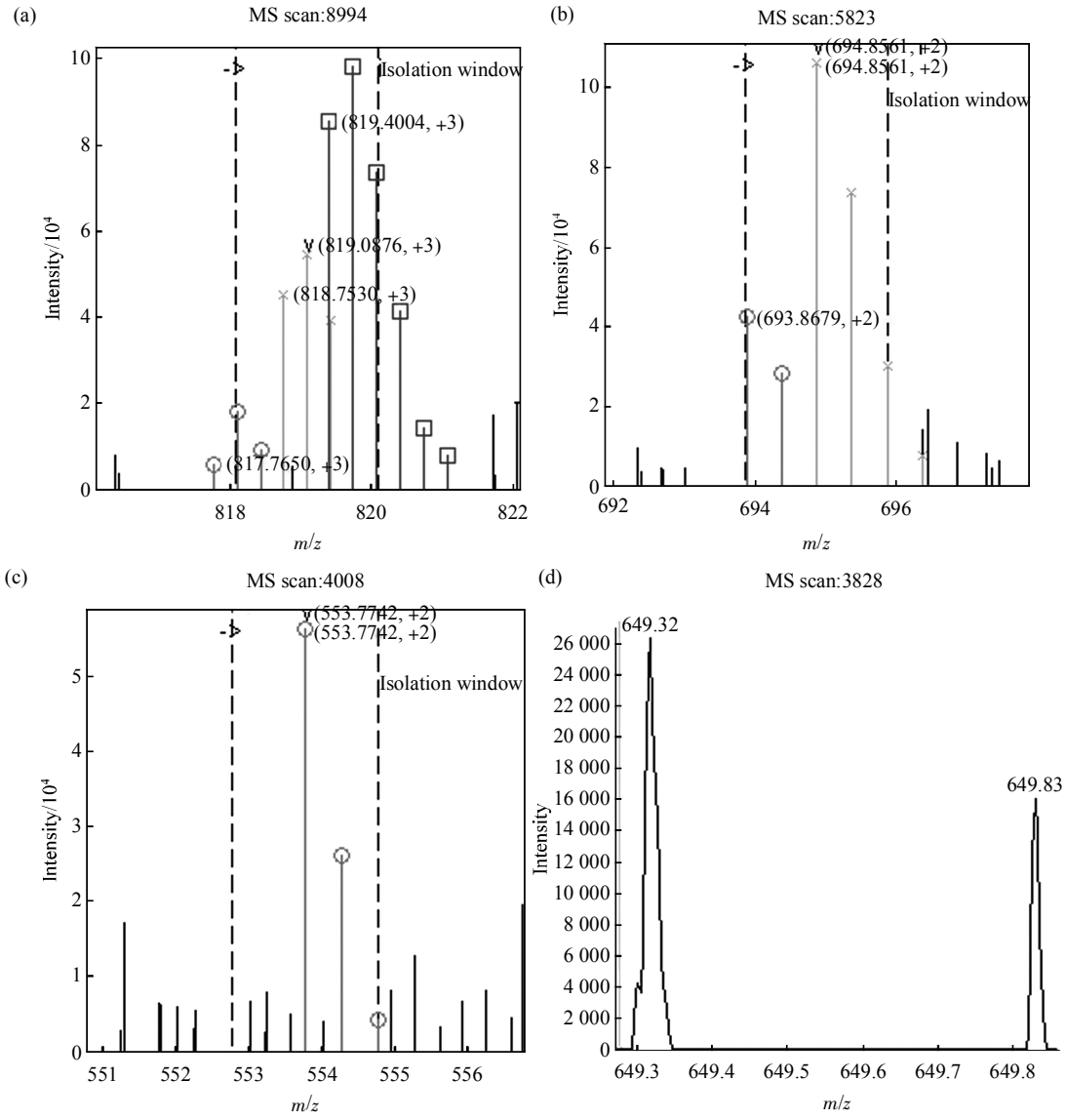

Fig. 5　Incorrect examples of MaxQuant

图 5　MaxQuant 判错的示例

(a) MaxQuant 只导出了第一个标 "x" 的峰. (b) MaxQuant 只导出了第二个标 "o" 的峰. (c) MaxQuant 导出了第一个标 "o" 的峰但是判为了+4 电荷. (d) 第一个峰被干扰了，导致中心化后的质荷比不准确.

2.2.5　本课题组的方法．尽管判断肽段单同位素峰的软件很多，但准确性上还有所欠缺．原因是这些软件通常假设一张二级谱只来源于一个肽段．从数据依赖的采集模式看，由于碎裂窗口较大，共洗脱的几个肽段被碎裂，这样很多二级谱实际上是混合谱，对应的母离子来自多个肽段．在一级谱上判断单同位素峰实际上比较复杂，碎裂中心的同位素峰簇被干扰，单同位素峰和电荷都有可能判错．所以必须要考虑碎裂窗口内的所有同位素峰簇，并且要考虑到它们会发生重叠的情况．THRASH 和 Hardklor 都考虑了重叠同位素峰簇．本文作者发现，根据平均氨基酸模型可以计算出：当肽段质量小于 1 800 Da 时，单同位素峰最高；当肽段质量在 1 800 Da 和 3 300 Da 之间时，第一同位素峰最高；当肽段质量大于 3 300 Da 时，第二同位素峰最高．利用最高同位素峰出现的位置和肽段质量的

关系可以初步判断是否有重叠，再用平均氨基酸模型得到的同位素峰强度分布和实验同位素峰强度分布计算相似度进一步确认是否有重叠．给出了可能的同位素峰簇之后，重构它们的色谱峰，根据当前同位素峰簇的强度和、同位素峰强度分布的相似度、色谱峰的相似度给可能的单同位素峰排序，输出前 5 个单同位素峰．这就是对应软件 pParse 判断单同位素峰的方法[52]．其优点是考虑共洗脱肽段的干扰，并利用平均氨基酸模型推出判断重叠同位素峰簇的方法．最后的测试结果是在 yeast 数据的标注集上，pParse 的准确率达到 99%, MaxQuant 的准确率是 88%, 而 Extract_MSn 的准确率只有 80%.

2.2.6 二级谱和数据库搜索的方法．除了上面这些判断肽段单同位素峰的方法和软件，还可以从二级谱和数据库搜索的角度来考虑单同位素峰质量校准．Venable 等[53]利用互补碎片离子获得母离子的质量．缺点是获得的母离子质量精度和二级谱是一个量级的，在低精度二级谱上不太适用．另外最后的结果受互补离子对的比例限制．不过在高精度二级谱上，这个方法可以作为一个参考．直接用 Da 量级的误差搜索数据库，损失了高精度过滤候选肽的能力．但可以考虑在质量偏离 1、2 等位置设置 ppm 量级的误差搜索数据库[54]．这个方法能利用高精度过滤候选肽的能力，但没法鉴定共洗脱肽段．

2.3 判断蛋白质和肽段单同位素峰的区别

和判断蛋白质大分子的单同位素峰相比，判断肽段的单同位素峰时情况有了变化．当分子质量大于 15 kDa 时，单同位素峰强度是最高同位素峰的万分之一，甚至可能更小[9]．单同位素峰和最高峰的强度相差 4 个或者更多的数量级，而质谱仪的强度检测范围有限，可能检测不到单同位素峰．蛋白质的分子质量较大，检测不到单同位素峰的情况较多．肽段的分子质量一般小于 4 kDa，通常可以检测到单同位素峰．这是判断蛋白质和肽段单同位素峰的第一个区别．质谱仪的质荷比检测范围有限，为了能检测到蛋白质，质谱仪在离子化的过程中让蛋白质带了很高的电荷，比如几十甚至上百．而质谱仪的分辨率有限，这时仅根据相邻峰的质荷比间隔难以判断蛋白质的电荷．肽段的分子质量小，只需带较小的电荷就可以检测，比如十电荷以内．目前质谱仪的分辨率足以把带这些电荷的同位素峰分开．所以根据相邻峰的质荷比间隔就可以判断肽段的电荷．这是判断蛋白质和肽段单同位素峰的第二个区别．蛋白质经过纯化、分离之后重叠同位素峰簇的情况会大大减少．肽段即使经过分离，也存在共洗脱的现象，从而导致重叠情况的发生．而发生重叠后，实验同位素峰分布和理论同位素峰分布的差距变大，会导致单同位素峰判断出错．这是判断蛋白质和肽段单同位素峰的第三个区别．可以选择不复杂的蛋白质，比如没有标记、修饰等．但是肽段没法选择，样品做过标记，或者修饰较多，这时直接使用平均氨基酸模型会带来误差，需要修正平均氨基酸模型．这是判断蛋白质和肽段单同位素峰的第四个区别．因为这四个区别，在判断肽段的单同位素峰时需要做相应的调整．

2.4 校准母离子质量的系统误差

判断了肽段单同位素峰即做了母离子的单同位素峰质量校准，第二个质量校准是校准母离子单同位素峰质量的系统误差．系统误差由质谱仪引起，可以从质谱仪入手，比如引入 lock mass，即把空气中已知质量的分子离子聚集起来，在扫描一级谱时引入这些空气分子离子，通过它们的质量偏差对一级谱的质量进行校准[55]．比较常见的是在把实验样品送入质谱仪前，先把标准样品送入质谱仪，根据标准样品的质量偏差对一级谱的质量进行校准．在 FT 类的仪器中质谱仪先测到的是频域信号，后经过变换变成时域信号，FTICR 和 Orbitrap 有各自的变换公式，可以先拿到频域信号的数据，学习和调整变换公式中的参数，获得参数后再变成时域信号[56]．这三个方法要求熟悉质谱仪，从源头上控制系统误差．DtaRefinery 的方法是先做预搜索，统计鉴定肽段的质量误差与保留时间、肽段强度、质量等的关系，做曲线拟合，获得系统误差的曲线，从母离子质量中减去系统误差从而进行校准[57]．MaxQuant 也做了类似的工作，只考虑了质量误差和保留时间、肽段质量的关系[58]，还利用肽段的电荷对做了进一步的质量校准[59]．pFind[60-61]的数据分析流程 Nezha 采取的方法是先做预搜索，统计质量偏差的分布，计算质量偏差的均值和标准差，从而给出误差窗口(中心不一定为 0)，根据每个质谱文件统计的误差窗口分布进行第二遍常规的搜索(http://159.226.41.114/nezha/login/)．后面三个方法的特点都是通过预搜索得到质量误差的分布，再做曲线拟合或计算窗口．通过这两大类方法就可以校准一级谱质量的系统误差．

3 总结和展望

本文综述了规模化蛋白质鉴定中母离子准确检测的工作，包括两大类：判断单同位素峰和系统误差校准．第一大类可以借鉴蛋白质水平的方法，包括电荷判断、单同位素峰判断和重叠同位素峰判断三个方面，同时在肽段水平判断单同位素峰有自己的特点．第二大类从质谱仪和后处理的角度综述了母离子质量的系统误差校准．第一大类是本文的重点．

上面从 THRASH 到 pParse 判断肽段单同位素峰的方法，都只是利用了一级谱的信息，包括同位素峰强度、色谱峰等．尽管利用的信息比较全面，但一级谱的固有缺点是无法确定肽段的身份，导致单同位素峰的确定有本质的困难．和一级谱相比，二级谱正好能克服一级谱的缺点，即能确定肽段的序列即身份，这样单同位素峰能确定下来．当然二级谱的鉴定依赖于母离子的质量．这样可以把一级谱的全面信息和二级谱的确定信息结合起来．思路如下：利用二级谱的母离子找到对应的碎裂窗口，把碎裂窗口内所有可能的同位素峰簇全部输出，每个同位素峰都作为可能的单同位素峰，这样原来的一张二级谱变成多张二级谱，碎片离子的信息是一样的，不同的是母离子的信息，把所有重新导出的二级谱送到搜索引擎(比如 pFind)去鉴定，对于有鉴定结果的母离子，返回一级谱查找同位素峰簇、重构色谱峰，计算同位素峰强度分布的相似度、色谱峰的相似度，再加上鉴定分数这三个特征来判断当前碎裂窗口最可能的单同位素峰．这个方法的优点是综合了一级谱和二级谱的优势．pParse 也使用了强度分布的相似度、色谱峰的相似度，第三个特征是当前同位素峰簇的强度和．而在这个新思路中第三个特征是肽谱匹配打分，因此这个方法比 pParse 的特征确定性更强，准确性更高．pParse 在碎裂窗口内的可能同位素峰簇非常多时容易出错，因为正确的单同位素峰可能排不到前 5，可能输出了错误的判断，漏掉了正确的判断．可以看出，这个方法将比 pParse 的准确率还高．

从上面的综述可以看出母离子的准确检测，已经有了比较全面和成熟的工作．不仅可以从质谱仪上做控制，从计算的角度也可以进一步提高单同位素峰质量的准确性和精度，从而为后续的蛋白质鉴定奠定坚实的基础．

致谢 感谢中国科学院计算技术研究所的付岩、王乐珩、王海鹏、张京芬等在算法、软件方面的讨论，军事医学科学院放射医学研究所、北京生命科学研究所等单位的大力支持．

参 考 文 献

[1] 马 洁, 吴松锋, 朱云平. 蛋白质组学中新蛋白质鉴定的研究方法和策略. 生物化学与生物物理进展, 2007, **34**(8): 791−799
Ma J, Wu S F, Zhu Y P. Prog Biochem Biophys, 2007, **34**(8): 791−799

[2] 李 宁, 吴松锋, 朱云平, 等. 鸟枪法蛋白质鉴定质量控制方法研究进展. 生物化学与生物物理进展, 2009, **36**(6): 668−675
Li N, Wu S F, Zhu Y P, *et al*. Prog Biochem Biophys, 2009, **36**(6): 668−675

[3] Marshall A G, Hendrickson C L, Jackson G S. Fourier transform ion cyclotron resonance mass spectrometry: a primer. Mass Spectrom Rev, 1998, **17**(1): 1−35

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

[5] Strittmatter E F, Rodriguez N, Smith R D. High mass measurement accuracy determination for proteomics using multivariate regression fitting: application to electrospray ionization time-of-flight mass spectrometry. Anal Chem, 2003, **75**(3): 460−468

[6] Petyuk V A, Jaitly N, Moore R J, *et al*. Elimination of systematic mass measurement errors in liquid chromatography-mass spectrometry based proteomics using regression models and a priori partial knowledge of the sample content. Anal Chem, 2008, **80**(3): 693−706

[7] Zubarev R, Mann M. On the proper use of mass accuracy in proteomics. Mol Cell Proteomics, 2007, **6**(3): 377−381

[8] Zubarev R A, Bondarenko P V. An A priori relationship between the average and monoisotopic masses of peptides and oligonucleotides. Rapid Commun Mass Spectrom, 1991, **5**(6): 276−277

[9] Senko M W, Beu S C, Mclafferty F W. Determination of monoisotopic masses and Ion populations for large biomolecules from resolved isotopic distributions. J Am Soc Mass Spectrom, 1995, **6**(4): 229−233

[10] Senko M W, Beu S C, Mclafferty F W. Automated assignment of charge states from resolved isotopic peaks for multiply charged ions. J Am Soc Mass Spectrom, 1995, **6**(1): 52−56

[11] Labowsky M, Whitehouse C, Fenn J B. Three-dimensional deconvolution of multiply charged spectra. Rapid Commun Mass Spectrom, 1993, **7**(1): 71−84

[12] Yergey J A. A general approach to calculating isotopic distributions for mass spectrometry. Int J Mass Spectrom Ion Phys, 1983, **52**(2): 337−349

[13] 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

[14] Zabrouskov V, Senko M W, Du Y, *et al*. New and automated MSn approaches for top-down identification of modified proteins. J Am Soc Mass Spectrom, 2005, **16**(12): 2027-2038

[15] Chen L, Sze S K, Yang H. Automated intensity descent algorithm for interpretation of complex high-resolution mass spectra. Anal Chem, 2006, **78**(14): 5006-5018

[16] Kaur P, O'Connor P B. Algorithms for automatic interpretation of high resolution mass spectra. J Am Soc Mass Spectrom, 2006, **17**(3): 459-468

[17] Barbarini N, Magni P. Accurate peak list extraction from proteomic mass spectra for identification and profiling studies. BMC Bioinformatics, 2010, **11**(1): 518

[18] Carvalho P C, Xu T, Han X, *et al*. YADA: a tool for taking the most out of high-resolution spectra. Bioinformatics, 2009, **25**(20): 2734-2736

[19] 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

[20] Jaitly N, Mayampurath A, Littlefield K, *et al*. Decon2LS: An open-source software package for automated processing and visualization of high resolution mass spectrometry data. BMC Bioinformatics, 2009, **10**(1): 87

[21] Gras R, Muller M, Gasteiger E, *et al*. Improving protein identification from peptide mass fingerprinting through a parameterized multi-level scoring algorithm and an optimized peak detection. Electrophoresis, 1999, **20**(18): 3535-3550

[22] Breen E J, Hopwood F G, Williams K L, *et al*. Automatic poisson peak harvesting for high throughput protein identification. Electrophoresis, 2000, **21**(11): 2243-2251

[23] Wehofsky M, Hoffmann R. Isotopic deconvolution of matrix-assisted laser desorption/ionization mass spectra for substance-class specific analysis of complex samples. Eur J Mass Spectrom, 2001, **7**(1): 39-46

[24] Wehofsky M, Hoffmann R. Automated deconvolution and deisotoping of electrospray mass spectra. J Mass Spectrom, 2002, **37**(2): 223-229

[25] Li X J, Yi E C, Kemp C J, *et al*. A software suite for the generation and comparison of peptide arrays from sets of data collected by liquid chromatography-mass spectrometry. Mol Cell Proteomics, 2005, **4**(9): 1328-1340

[26] Noy K, Fasulo D. Improved model-based, platform-independent feature extraction for mass spectrometry. Bioinformatics, 2007, **23**(19): 2528-2535

[27] Bellew M, Coram M, Fitzgibbon M, *et al*. A suite of algorithms for the comprehensive analysis of complex protein mixtures using high-resolution LC-MS. Bioinformatics, 2006, **22**(15): 1902-1909

[28] Mortensen P, Gouw J W, Olsen J V, *et al*. MSQuant, an open source platform for mass spectrometry-based quantitative proteomics. J Proteome Res, 2010, **9**(1): 393-403

[29] Hoopmann M R, Finney G L, MacCoss M J. High-speed data reduction, feature detection, and MS/MS spectrum quality assessment of shotgun proteomics data sets using high-resolution mass spectrometry. Anal Chem, 2007, **79**(15): 5620-5632

[30] Fernandez-de-Cossio J, Gonzalez L J, Satomi Y, *et al*. Automated interpretation of mass spectra of complex mixtures by matching of isotope peak distributions. Rapid Commun Mass Spectrom, 2004, **18**(20): 2465-2472

[31] Zhang X, Hines W, Adamec J, *et al*. An automated method for the analysis of stable isotope labeling data in proteomics. J Am Soc Mass Spectrom, 2005, **16**(7): 1181-1191

[32] Mason C J, Therneau T M, Eckel-Passow J E, *et al*. A method for automatically interpreting mass spectra of ^{18}O-labeled isotopic clusters. Mol Cell Proteomics, 2007, **6**(2): 305-318

[33] Roussis S G, Proulx R. Reduction of chemical formulas from the isotopic peak distributions of high-resolution mass spectra. Anal Chem, 2003, **75**(6): 1470-1482

[34] Bielow C, Ruzek S, Huber C G, *et al*. Optimal decharging and clustering of charge ladders generated in ESI-MS. J Proteome Res, 2010, **9**(5): 2688-2695

[35] Samuelsson J, Dalevi D, Levander F, *et al*. Modular, scriptable and automated analysis tools for high-throughput peptide mass fingerprinting. Bioinformatics, 2004, **20**(18): 3628-3635

[36] Du P, Angeletti R H. Automatic deconvolution of isotope-resolved mass spectra using variable selection and quantized peptide mass distribution. Anal Chem, 2006, **78**(10): 3385-3392

[37] Renard B Y, Kirchner M, Steen H, *et al*. NITPICK: peak identification for mass spectrometry data. BMC Bioinformatics, 2008, **9**(1): 355

[38] McIlwain S, Page D, Huttlin E L, *et al*. Using dynamic programming to create isotopic distribution maps from mass spectra. Bioinformatics, 2007, **23**(13): i328-336

[39] Sun Y, Zhang J, Braga-Neto U, *et al*. BPDA - a Bayesian peptide detection algorithm for mass spectrometry. BMC Bioinformatics, 2010, **11**(1): 490

[40] Mayampurath A M, Jaitly N, Purvine S O, *et al*. DeconMSn: a software tool for accurate parent ion monoisotopic mass determination for tandem mass spectra. Bioinformatics, 2008, **24**(7): 1021-1023

[41] Shin B, Jung H J, Hyung S W, *et al*. Postexperiment monoisotopic mass filtering and refinement (PE-MMR) of tandem mass spectrometric data increases accuracy of peptide identification in LC/MS/MS. Mol Cell Proteomics, 2008, **7**(6): 1124-1134

[42] Jung H J, Purvine S O, Kim H, *et al*. Integrated post-experiment monoisotopic mass refinement: an integrated approach to accurately assign monoisotopic precursor masses to tandem mass spectrometric data. Anal Chem, 2010, **82**(20): 8510-8518

[43] Scherl A, Tsai Y S, Shaffer S A, *et al*. Increasing information from shotgun proteomic data by accounting for misassigned precursor ion masses. Proteomics, 2008, **8**(14): 2791-2797

[44] Park K, Yoon J Y, Lee S, *et al*. Isotopic peak intensity ratio based algorithm for determination of isotopic clusters and monoisotopic masses of polypeptides from high-resolution mass spectrometric data. Anal Chem, 2008, **80**(19): 7294-7303

[45] Cox J, Mann M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat Biotechnol, 2008, **26**(12): 1367−1372

[46] Sturm M, Bertsch A, Gropl C, *et al*. OpenMS - an open-source software framework for mass spectrometry. BMC Bioinformatics, 2008, **9**(1): 163

[47] Monroe M E, Tolic N, Jaitly N, *et al*. VIPER: an advanced software package to support high-throughput LC-MS peptide identification. Bioinformatics, 2007, **23**(15): 2021−2023

[48] Mueller L N, Rinner O, Schmidt A, *et al*. SuperHirn - a novel tool for high resolution LC-MS-based peptide/protein profiling. Proteomics, 2007, **7**(19): 3470−3480

[49] Leptos K C, Sarracino D A, Jaffe J D, *et al*. MapQuant: open-source software for large-scale protein quantification. Proteomics, 2006, **6**(16): 1770−1782

[50] Katajamaa M, Miettinen J, Oresic M. MZmine: toolbox for processing and visualization of mass spectrometry based molecular profile data. Bioinformatics, 2006, **22**(5): 634−636

[51] McDonald W H, Tabb D L, Sadygov R G, *et al*. MS1, MS2, and SQT-three unified, compact, and easily parsed file formats for the storage of shotgun proteomic spectra and identifications. Rapid Commun Mass Spectrom, 2004, **18**(18): 2162−2168

[52] Yuan Z F, Liu C, Wang H P, *et al*. pParse: a method for accurate determination of monoisotopic peaks in high-resolution mass spectra. Proteomics, 2012, **12**(2): 226−235

[53] Venable J D, Xu T, Cociorva D, *et al*. Cross-correlation algorithm for calculation of peptide molecular weight from tandem mass spectra. Anal Chem, 2006, **78**(6): 1921−1929

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

[55] Olsen J V, de Godoy L M, Li G, *et al*. Parts per million mass accuracy on an Orbitrap mass spectrometer *via* lock mass injection into a C-trap. Mol Cell Proteomics, 2005, **4**(12): 2010−2021

[56] Haas W, Faherty B K, Gerber S A, *et al*. Optimization and use of peptide mass measurement accuracy in shotgun proteomics. Mol Cell Proteomics, 2006, **5**(7): 1326−1337

[57] Petyuk V A, Mayampurath A M, Monroe M E, *et al*. DtaRefinery, a software tool for elimination of systematic errors from parent ion mass measurements in tandem mass spectra data sets. Mol Cell Proteomics, 2010, **9**(3): 486−496

[58] Cox J, Michalski A, Mann M. Software lock mass by two-dimensional minimization of peptide mass errors. J Am Soc Mass Spectrom, 2011, **22**(8): 1373−1380

[59] Cox J, Mann M. Computational principles of determining and improving mass precision and accuracy for proteome measurements in an Orbitrap. J Am Soc Mass Spectrom, 2009, **20**(8): 1477−1485

[60] Fu Y, Yang Q, Sun R, *et al*. Exploiting the kernel trick to correlate fragment ions for peptide identification *via* tandem mass spectrometry. Bioinformatics, 2004, **20**(12): 1948−1954

[61] Wang L H, Li D Q, Fu Y, *et al*. pFind 2.0: a software package for peptide and protein identification *via* tandem mass spectrometry. Rapid Commun Mass Spectrom, 2007, **21**(18): 2985−2991

Accurate Determination of Precursor Ions for Peptides in Large-scale Protein Identification*

YUAN Zuo-Fei[1,2], WU Long[1,2], LIU Chao[1,2], CHI Hao[1,2], FAN Sheng-Bo[1,2], ZHANG Kun[1,2], ZENG Wen-Feng[1,2], SUN Rui-Xiang[1], HE Si-Min[1]**

([1] *Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, The Chinese Academy of Sciences, Beijing* 100190, China;
[2] *University of Chinese Academy of Sciences, Beijing* 100049, China)

Abstract One basic research for proteomics is protein identification. For large-scale protein identification, shotgun techniques are usually applied, *i.e.*, some digested peptides (precursor ions) are chosen to fragment and produce tandem mass spectra, which can be identified by database search, and then proteins can be inferred from the identified peptides. In the identification, precursor mass is a key parameter. Whether the correct peptide is within the peptide candidates depends on whether the precursor mass is the monoisotopic mass. The number of peptide candidates depends on the accuracy of the precursor mass. This research investigates the accurate determination of precursors in terms of monoisotopic peak determination and systematic error elimination. There are some techniques of monoisotopic peak determination on protein level, including charge state determination, monoisotopic peak determination and overlapping cluster determination. Some of them can be used to the monoisotopic peak determination of precursors. Meanwhile, there are some well-known methods for systematic error elimination. The two kinds of techniques for accurate precursor determination can help increase the large-scale protein identification rate.

Key words proteomics, protein identification, precursor ions, monoisotopic peaks, systematic error
DOI: 10.3724/SP.J.1206.2012.00167

*This work was supported by grants from the National Basic Research Program of China (2010CB912701, 2012CB910602), The CAS Knowledge Innovation Program(KGGX1-YW-13), The National Natural Science Foundation of China(30900262) and Hi-Tech Research and Development Program of China (2007AA02Z315, 2008AA02Z309).

**Corresponding author.
Tel: 86-10-62601016, E-mail: smhe@ict.ac.cn

Received: April 1, 2012 Accepted: June 5, 2012