PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2011, 38(6): 506~518
www.pibb.ac.cn

蛋白质质谱分析的无标记定量算法研究进展 *

张 伟[1] 张纪阳[1] 刘 辉[1] 孙汉昌[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 作为发现疾病相关生物标志物的重要途径，定量研究已成为蛋白质组学的热点问题．随着实验方法的发展和改进，定量数据处理算法也在不断更新和完善．将现有的无标记定量方法归纳为需要 / 不需要鉴定结果两类方法，分析比较了两类方法的异同及优缺点，详细讨论了所涉及的主要算法，总结了一些常用的无标记定量软件及对应的网络资源．展望了无标记定量数据分析的未来研究方向．

关键词 定量蛋白质组，质谱，无标记定量，定量算法，统计学分析
学科分类号 Q51，TP391 **DOI**: 10.3724/SP.J.1206.2010.00560

细胞蛋白质组是一个高度动态的系统，其细微变化就可以引起某些重要通路的激活或抑制，进而可能导致细胞的变异、分裂，甚至凋亡．因此，蛋白质丰度的变化也是许多疾病(如癌症、心血管疾病等)发病进程的重要标志，研究病变细胞相对于正常细胞中蛋白质表达丰度的变化，发现标志物对疾病的早期诊断具有重要意义．要完成这一工作，就需要对细胞中表达的全套蛋白质进行定量分析．正因为如此，定量蛋白质组已成为蛋白质组学中的一个重要分支．

质谱分析技术是实现大规模、高通量蛋白质定量的主要方法．尽管现阶段基于质谱的蛋白质定量主要局限于相对定量研究，但是由于质谱仪器精度和灵敏度的不断提高，基于质谱的蛋白质组定量越来越受青睐[1]．从实验设计上来看，基于质谱的定量分析包括稳定同位素标记(stable isotopic labeling)和无标记(label-free)两种方法[2]．其中，稳定同位素标记法通过代谢、化学标记等方法在肽段上引入质量标签，在同一次实验中分析不同标记的混合样本，同时得到不同样本中肽段 / 蛋白质的响应信号，标记方法定量的精度较高，但是其实验复杂、昂贵，动态范围和覆盖率受到标记方法的限制[3]．相比之下，无标记定量对不同状态下的样本单独进行质谱分析，虽然对实验的可重复性要求较高，但克服了上述稳定同位素标记定量的技术局限，应用范围越来越广．经过近几年的发展，无标记定量分析已发展出几套成熟的实验分析策略和相应的计算流程，针对不同的实验仪器和数据，定量算法和数据分析软件也层出不穷．

本文首先总结了无标记定量数据分析的几种典型流程，比较了它们的差异，总结了各种方法的优缺点．在此基础上，详细分析了无标记定量数据分析流程中的关键算法，着重论述了定量算法的现状、存在的问题．最后总结了一些常用的无标记定量数据分析软件及相应的网络资源．

1 无标记定量及其典型计算流程

在定量蛋白质组学中，“鸟枪法”是常用的实验策略，图 1 给出了其实验流程．通常情况下，由于物理化学特性的不同，由蛋白质酶切得到的肽段混合物会在不同时间流出液相色谱，进入质谱仪进行质谱分析，得到包含肽段定量信息的一级图谱(MS spectrum)和包含肽段序列信息的二级图谱(MS/MS spectrum)．

* 国家青年自然科学基金(31000587)和蛋白质组学国家重点实验室开放课题(SKLP-O201004)资助项目．
** 通讯联系人．
Tel: 0731-84576311, E-mail: xhwei65@nudt.edu.cn
收稿日期：2010-11-01，接受日期：2011-01-06

无标记定量主要有 LC-MS 和 LC-MS/MS 两种实验策略[4]，其主要差别在于是否利用串联质谱分析来鉴定肽段和蛋白质．两种实验策略在数据分析流程上有很大不同．因此，本文将无标记定量方法分成无需 / 需要鉴定结果的定量方法，其计算流程分别对应于图 1 中的流程一和流程二．

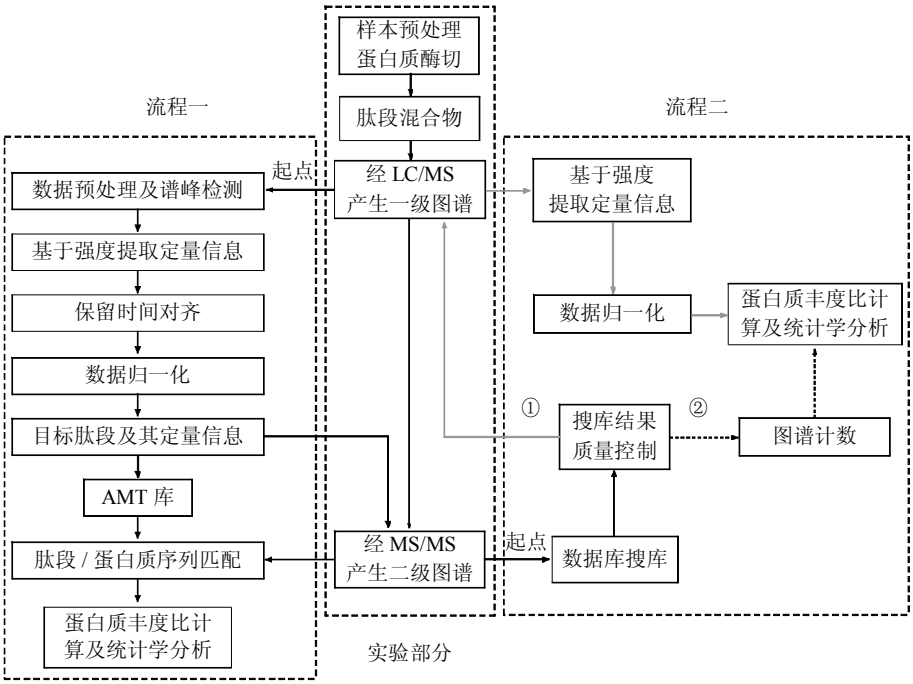

Fig. 1　Workflow of label-free proteomics quantification analysis

图 1　无标记蛋白质组定量分析的典型计算流程

无需鉴定结果的定量方法以一级图谱数据为处理对象，其定量数据处理主要由以下 6 步完成：a. 数据预处理及谱峰检测(peak detection)．主要目的是从含有大量噪声的单张一级图谱中提取真实的肽段信号峰．b．基于信号强度(intensity)提取肽段定量信息．在保留时间(retention time, RT)轴上，构建肽段的离子流色谱峰(extracted ion chromatography, XIC)，并根据 XIC 计算出肽段的丰度表征．c．保留时间对齐(RT alignment)．目的是为了消除不同实验中同一肽段的色谱保留时间偏差．d．数据归一化(normalization)．消除不同实验之间肽段信号强度的系统误差．e．肽段 / 蛋白质序列匹配．无序列信息的目标肽段可以通过精确质量时间标签(accurate mass and time, AMT)进行数据库搜索[5]或通过靶标式 LC-MS/MS 分析[6]匹配到肽段 / 蛋白质序列．f．蛋白质丰度比计算及统计学分析．由肽段的定量值推断出对应蛋白质的丰度比，然后通过统计学分析找出显著性差异表达的蛋白质，从而确定候选生物标志物．值得注意的是，在临床诊断中可能不需要肽段和蛋白质的序列信息，而是构建特定生物样品的质谱分析特征矩阵，利用数据特征直接刻画或者表征样品[7-9]．

需要鉴定结果的定量方法是针对 LC-MS/MS 策略的实验数据处理方法，其数据处理步骤包括：a．数据库搜索及结果质量控制．利用二级图谱，通过数据库搜索和结果质量控制，得到高可信度的肽段和蛋白质的鉴定结果．b．定量信息提取．有两种不同方法——信号强度法和图谱计数(spectral counting)法，分别对应图 1 中流程二的①和②．方法①利用肽段的鉴定信息返回到一级图谱中提取肽段的 XIC，并根据 XIC 计算肽段的丰度表征；方法②则把蛋白质中肽段的鉴定图谱总数作为定量指标，只能定量蛋白质．c．蛋白质丰度比计算及统计学分析．

由于采用了不同的实验策略，两种计算流程各有优缺点．流程一采用了 LC-MS 实验策略，直接从一级图谱中检测肽段特征并提取定量信息．由于不需要选择母离子，在合适的结果过滤规则下，流程一可以定量更多的肽段，对低丰度肽段的定量有利，但是存在假阳性率较高、多肽段重叠的情况[10]，并且定量算法比较复杂、运算时间很长．而流程二则可以利用肽段和蛋白质的鉴定结果完成定量，假阳性率很低．但由于采样效应的限制，肽段覆盖率较低[11]，并且大部分是高丰度肽段，而许多重要差异表达的生物标志物往往丰度较低[12]，这就不利于生物标志物的发现．目前为止，还没有相关文献对这两种数据处理流程的优劣进行系统评估．

2　定量算法

无标记定量的两种典型计算流程采用的实验策略不同，导致数据处理步骤有很大差异．a．流程一采用谱峰检测算法确定定量对象，而流程二的定量对象则通过数据库搜索和结果质量控制获取，其中数据库搜索鉴定一般由商业化的软件完成，例如 SEQUEST[13]和 Mascot[14]，结果质量控制也有比较成熟的方法和软件[15-18]．b．图谱计数法是流程二特有的定量方法．c．保留时间对齐是流程一必不可少的数据处理步骤，而流程二则不需要．d．在推断蛋白质丰度比之前，流程一需要匹配出肽段 / 蛋白质的序列．尽管如此，两种计算流程具有相同的数据归一化、蛋白质丰度比推算以及统计学分析步骤．下面对两种数据处理流程涉及的主要算法进行论述．

2.1　数据预处理及谱峰检测

数据预处理及谱峰检测是流程一的基础，其主要目的是从含有大量噪声的一级图谱中提取肽段信号峰．与二级图谱相比，一级图谱包含了所有检测到肽段的信息，但是其中只有很小一部分质谱信号属于肽段信号，其余为随机噪声、化学噪声等干扰信号．因此，准确快速地提取肽段信号峰至关重要．数据预处理和谱峰检测有很多可选的算法，针对不同的质谱数据，处理算法也不尽相同．

对于低精度的质谱数据，目前对 MADLI/SEDLI 实验数据的谱峰检测讨论较多[19-25]．图 2 给出了数据预处理及谱峰检测处理过程的一个具体例子，包含了如下 3 个步骤[24]：a．噪声滤波．主要为了去除图谱中的随机噪声，有移动平均滤波、Savitzky-Golay 滤波、高斯滤波、连续或离散小波变换、Hilbert-Huang 变换等算法．b．基线去除．估计并去除图谱中的基线，其算法包括单调局部最小、线

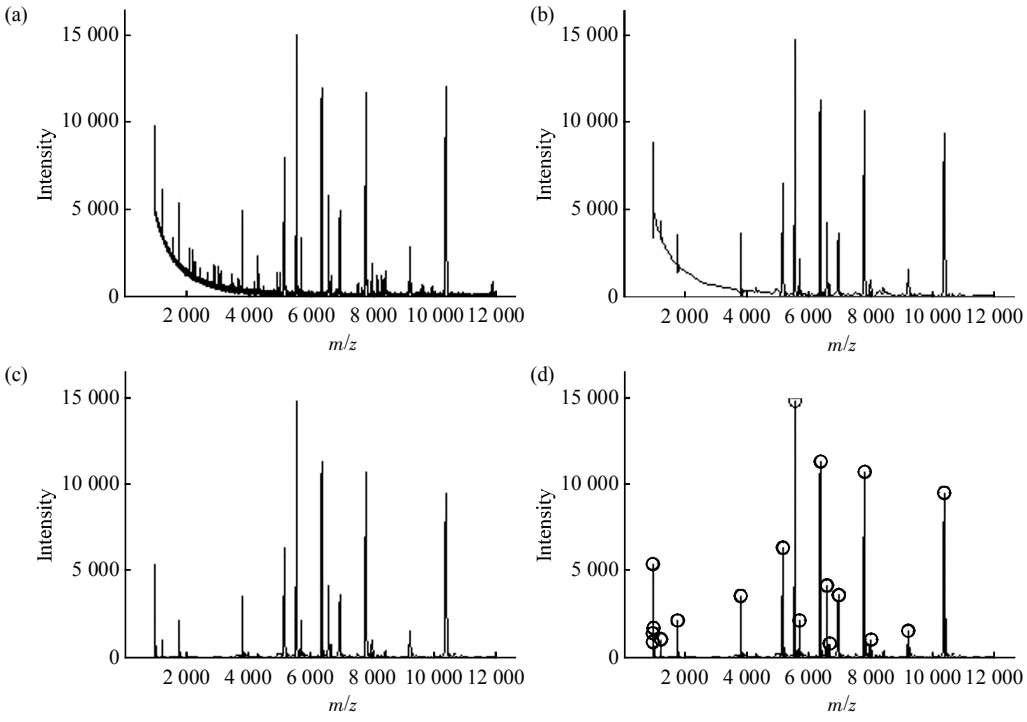

Fig. 2　An example of the data preprocessing and peak detection for low-resolution MS data

图 2　低精度质谱数据的预处理及谱峰检测过程示例

(a)原始质谱图. (b)噪声滤波以后的质谱图. (c)滤波和基线去除以后的质谱图. (d)谱峰检测的最终结果，真实的肽段峰用圆圈标记.

性插值、移动平均最小值、连续小波变换等. c. 峰识别. 主要从噪声和基线去除后的数据中识别出肽段信号峰，峰识别的准则有信噪比、峰强度阈值、局部最大值、峰宽、峰形等. 上述 3 个步骤都有很多算法可供选择，并且不同算法组合的性能差别很大. Cruz-Marcelo 等[23]和 Yang 等[24]对之前的处理低精度 MADLI/SEDLI 实验数据的谱峰检测算法进行了综述与评估，一致认为 Du 等[21]的基于连续小波变换的谱峰检测算法整体效果最好. 最近，Wu 等[25]基于 Hilbert-Huang 变换提出了一种新的峰检测算法，性能优于以往算法.

对于高精度的质谱数据，其谱峰检测与低精度数据类似，但可以利用高精度数据的特点过滤噪声信号峰. 本文将此类谱峰检测算法归纳为两大类：

a. 利用肽段的天然同位素分布过滤噪声信号[7,12]. 由于噪声信号通常不会显示为具有一定保留时间的多电荷同位素分布峰，所以这一规则可以保证在去除噪声干扰信号的同时，识别强度较低的肽段离子信号. 另外，根据同位素分布模式可以判别出肽段离子的电荷状态. 例如，Bellew 等[7]根据质荷比估计了肽段的天然同位素分布，通过打分函数计算了观测的同位素分布与天然同位素分布之间的相似度，设定相似度阈值，过滤掉阈值以下的噪声信号峰. 需要指出的是，若两个同位素分布模式部分重叠时，这类方法可能失效，并且估计的天然同位素分布与实际的分布存在偏差.

b. 利用 TOF 类数据中化学噪声的特点过滤噪声信号[26-27]. 高精度的 TOF 类质谱仪(如 MALDI-IT-TOF、prOTOF 等质谱仪)在具有高灵敏度的同时，把化学噪声也放大化了，其峰形与肽段信号峰类似. 此外，这些化学噪声峰的位置以 1Da 左右频率在 m/z 轴上呈现周期性，并且化学噪声峰的强度在局部范围内有波动，但整体上随着质荷比的增大而缓慢减少. 例如，McLerran 等[26]根据化学噪声峰的上述特点，通过估计其位置和强度的分布过滤掉化学噪声峰，Zhang 等[27]则利用正弦信号对化学噪声进行拟合，从而将其去除. 这类算法在去除化学噪声的同时，也丢失了那些强度低于或接近于化学噪声的肽段信号峰.

相比于低精度的质谱数据，处理高精度质谱数据的谱峰检测算法研究还较少，如何利用高精度数据的特点快速准确地检测出肽段信号峰是今后谱峰检测的研究重点.

2.2　定量信息提取

定量信息提取是定量数据处理中的基本步骤，在很大程度上决定了定量结果的精度，主要完成计算肽段或蛋白质定量指标的工作. 目前的定量信息提取方法主要有两大类：a. 信号强度法；b. 图谱计数法.

2.2.1　信号强度法.

信号强度法提取肽段定量信息的示例如图 3 所示，主要包括从一级图谱中解析肽段信号、构建肽

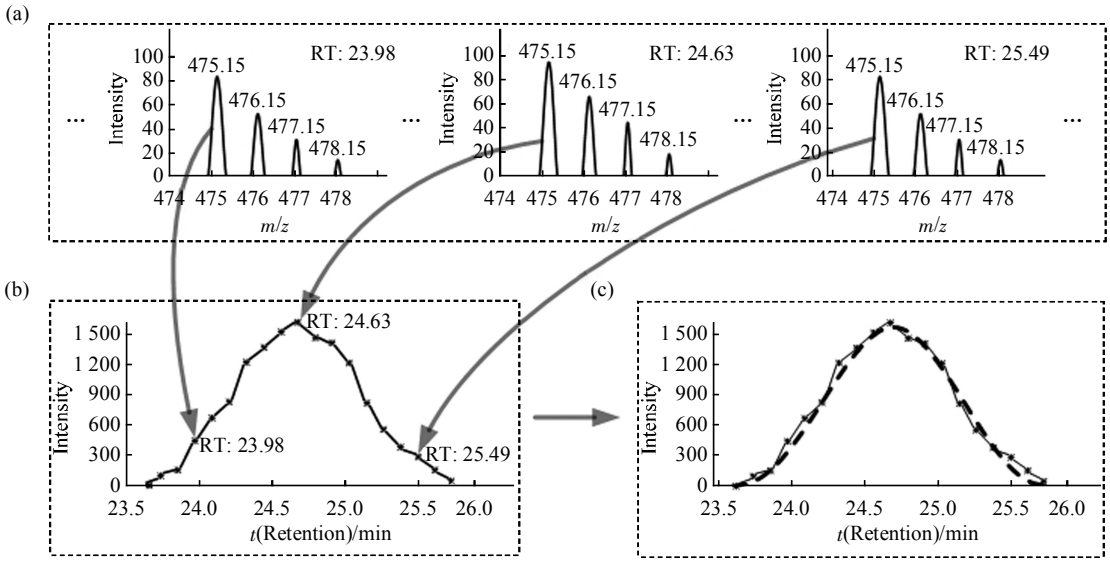

Fig. 3　An example of extracting quantitative information for the peptide with m/z=475.15

图 3　质荷比为 475.15 肽段的定量信息提取流程示例

(a)包含质荷比为 475.15 的肽段信息的一级图谱. (b)沿保留时间构建的肽段离子流色谱峰. (c)处理后的 XIC，以图中的虚线表示.

段沿保留时间展开的 XIC、处理 XIC 并计算肽段定量指标．流程一和流程二的此类定量信息提取算法类似，但是在提取定量信息之前，前者需要采用质荷比误差匹配原则、聚类等峰对齐算法[27-29]，识别出不同图谱中相同的肽段信号峰．

目前，这类定量信息提取方法有很多[7, 12, 30-34]，其区别主要表现在如下 5 个方面：a．去噪方法．解析肽段信号之前，是否对一级图谱去噪处理，其中去噪方法有小波去噪、滑动平均去噪、Savitzky-Golay 滤波等．b．肽段信号的图谱解析．采用肽段信号峰的峰值、峰内信号强度加和、峰平滑后的面积以及峰拟合后的面积等方法来完成从一级图谱中解析肽段信号．c．同位素峰．可以使用单一同位素峰、信号强度最高的同位素峰或前三个同位素峰来提取肽段的定量信息．d．XIC 的处理方法．使用小波去噪、平滑去噪、正则化、连续性截断等方法处理 XIC，或者对 XIC 不作处理．e．计算定量指标．把处理后 XIC 的峰值、峰内信号强度加和或者峰面积作为肽段的定量值．对于无需鉴定结果的定量方法，Bellew 等[7]对原始图谱进行小波去噪处理后检测肽段信号峰，同时采用谱峰的最大值作为肽段在某时刻的信号，最后使用 XIC 的峰值作为肽段的定量值，Li 等[12]考虑了肽段的前 3 个同位素峰，使用谱峰内信号的加和来构建 XIC，肽段的定量值为 Savitzky-Golay 平滑后的 XIC 峰面积．对于需要鉴定结果的定量方法，Tsou 等[32]利用聚类方法确定了肽段的前 3 个同位素峰的质荷比和保留时间范围，采用质荷比范围内信号的加和解析肽段信号，构建保留时间范围内的 XIC，最后的定量值为 B- 样条平滑后 XIC 的峰面积，Yang 等[33]利用鉴定肽段的前 3 个同位素峰在谱峰中的峰值来构建 XIC，并通过正则化方法处理了 XIC．

定量信息提取方法在上述 5 个方面的不同组合必然会导致不同的定量结果，评估这 5 个方面对定量结果的影响，并且选出一组或几组最优的算法组合是一件有意义的事情，但是，到目前为止还没有相关文献报道此类工作．

2.2.2 图谱计数法．

根据蛋白质丰度越高、对应肽段被鉴定的概率就越大的原理，图谱计数法不需要各种复杂的数据处理步骤，只需统计肽段的鉴定图谱数，把蛋白质中肽段的鉴定图谱总数作为定量指标．图谱计数法在 2004 年由 Liu 等[35]提出，他们通过分析标准蛋白质的质谱数据，揭示了在超过两个数量级的范围内，蛋白质的鉴定图谱总数与其浓度呈线性关系．

由于概念简单、运算速度快等特点，图谱计数方法吸引了不少学者的关注．为了进一步提高这类方法的实用性，现已发展了多种校正的图谱计数方法．Ishihama 等[36]指数修正了蛋白质丰度指标(emPAI)．Zybailov 等[37]利用蛋白质的序列长度校正了蛋白质的图谱总数，提出了 NSAF(normalized spectral abundance factor)指标．Lu 等[38]预测了蛋白质的检测效率，把使用检测效率校正后的蛋白质图谱总数作为最终的定量指标(APEX)．Sun 等[39]则通过蛋白质的鉴定概率来校正图谱数．值得注意的是，2010 年，Griffin 等[40]将肽段碎片离子的信号强度与图谱计数结合起来，得到了一个新的定量指标(SI_N)，与 NASF 指标[37]、基于信号强度法的 AUC 方法[34]等相比，SI_N 指标在可重复性和定量准确性方面整体较好，但是文献中使用的数据集为低精度的 LTQ 质谱数据．

信号强度法和图谱计数法都是常用的定量方法，不少学者系统评估了这两种定量方法的优劣．针对 LC-MS/MS 策略的实验数据，Old 等[34]的研究表明，图谱计数方法在检测显著差异表达的蛋白质方面更加灵敏，但是对于鉴定图谱总数很少的蛋白质，这类方法往往会过度估计其丰度比，而信号强度法能够更加准确地估计蛋白质的丰度比，且不受鉴定图谱数的影响，但数据处理流程相对复杂，运算速度较慢．Zybailov 等[41]认为图谱计数法具有更好的可重复性．Asara 等[42]的结果表明，信号强度法可估计的定量结果的动态范围更大．Xia 等[43]发现，当估计的蛋白质丰度比与真实的丰度比相关时，图谱计数法估计的丰度比比峰强度法准确．上述评估都是在中 - 低精度的质谱数据上进行的．2010 年，对 FT-LTQ 高精度数据，Grossmann 等[44]实现了信号强度法定量算法——T3PQ，并与 emPAI[36]和 APEX[38]定量指标进行了比较，结果表明，不论在动态范围方面，还是在定量准确性和可重复性方面，信号强度法要优于图谱计数法．尽管如此，整合两种方法的定量结果可能是提高定量算法整体性能的有效途径．

2.3 保留时间对齐

保留时间对齐的主要目的是消除不同实验中同一肽段的色谱保留时间偏差．要比较不同状态下肽

段 / 蛋白质的表达差异，就必须辨别出不同实验中的相同肽段．需要鉴定结果的定量算法可以根据序列信息来辨别相同肽段，而无需鉴定结果的定量算法则是通过设置质荷比窗口和色谱保留时间窗口来实现．虽然不同实验中的同一肽段在质荷比轴上产生的偏差很小，但是在保留时间轴上却会发生很大偏移，所以实现不同实验间保留时间的对齐是精确定量的关键．2008 年，Vandenbogaert 等[45]对保留时间对齐算法进行了全面综述．总的来说，所有的保留时间对齐方法可以归纳为两大类：a．特征数据(Peak-based Alignment)法；b. 谱数据(Profile-based Alignment)法．

特征数据法使用谱峰检测提取的肽段信息实现对齐．谱峰检测可以把具有几百万个数据点的图谱缩减到只有几百或者几千个肽段信号峰的特征图谱，显著降低了计算复杂度．这类方法又有两种思路：一是利用两次实验中能明显确定为同一肽段信号的保留时间，通过拟合、回归等方法，建立两次实验间保留时间的线性或者非线性模型对应关系[7, 46]；另一种方法是把两次实验中所有可能的特征匹配关系列出，计算所有匹配的相似性，然后利用动态规划、模拟退火等优化方法计算出相似性最高的最优匹配[47]．前一种方法计算较简单，但关键是要确定出一小批不同实验中相同的肽段信号；后一种方法不需要确定出相同的肽段信号，但不同的相似性函数对算法性能影响较大．

谱数据法利用未经处理的原始质谱数据实现对齐．与特征数据法相比，庞大的数据量对算法和计算平台性能的要求明显更高，但是可以充分利用原始图谱中的许多有用信息．谱数据法将会成为未来保留时间对齐方法的研究热点[45]．这类方法的基本思想是：试图找到一种对齐模式，使得所有 MS 图谱与参考图谱的强度之间的整体差异最小．其中，Eilers 等[48]从原始质谱数据中提取总离子流色谱峰(total ion chromagrams, TIC)，利用 PTW(parametric time warping)算法对齐 TIC；Christin 等[49]则改进了 PTW 方法．这类对齐 TIC 的方法仍然丢弃了原始质谱数据中的一些有用信息．针对原始质谱数据，Kong 等[50]提出了一个新的基于 Bayesian 方法的保留时间对齐算法．

2.4 数据归一化

数据归一化的主要目的是消除不同实验间肽段信号的系统误差．在质谱实验中，由于不同的离子化效率、图谱采样效应等原因，即便是相同实验中浓度相等的不同肽段，或者是不同实验中浓度相等的同一肽段，其信号强度也可能存在很大偏差．因此，为了获得更加准确的定量结果，对肽段信号的归一化处理是十分必要的．数据归一化方法可分为两大类，第一类是在实验样本中加入“内标”或“外标”标准蛋白质，构建归一化标准曲线，第二类是基于统计学模型的归一化方法．前一类方法虽然高效，但是样本处理技术的复杂性限制了其使用，所以目前大都采用后一类方法．这类方法是在处理 DNA 微阵列数据时引进的[51]，大都被直接应用或间接推广到基于质谱分析的蛋白质定量数据中，其中包括了全局归一化、线性回归归一化、局部回归归一化、分位数归一化、LOWESS(locally weighted scatter plot smoothing)等．针对 LC-FTICR 质谱数据，Callister 等[52]评估了 4 种数据归一化方法，结果表明线性回归归一化方法在大部分情况下性能最好，并同时建议对不同类型的质谱数据需要选择不同的归一化方法．Kultima 等[53]则分析了 LC-QTOF 和 LC-LTQ 质谱数据，提出了一种结合实验次序的线性回归归一化方法，与其他 9 种常用的归一化方法相比，该方法的归一化效果最好．

2.5 蛋白质丰度比计算

蛋白质丰度比计算的主要目的是根据肽段的定量值推断出对应蛋白质的丰度比．除图谱计数法外，流程一和流程二都是肽段水平的定量，而定量分析的主要目的是从各组实验数据中找出显著性差异表达的蛋白质，所以蛋白质丰度比计算至关重要．

2010 年，Carrillo 等[54]评估了 6 种蛋白质丰度比计算方法，分别是肽段定量比值的平均值、肽段定量值之和的比值、Libra 比值、线性回归、主成分分析和总体最小二乘法，结果表明，使用肽段定量值之和的比值作为蛋白质丰度比的效果最好．但是像多数分定量方法一样，该研究并没有考虑蛋白质丰度比推算中两个重要的问题：a．数据缺失问题，即肽段的定量值在某些实验中存在，而在另一些实验中没有记录．b．共享肽段在不同蛋白质间的丰度分配问题，共享肽段即是匹配多个蛋白质的肽段．

数据缺失的肽段主要包括：a．在质谱数据中存在但未被鉴定的肽段；b．实验中未被仪器检测到的肽段．产生第二种数据缺失肽段的原因可能是

肽段的丰度低于仪器的检测水平，或者是离子化失效、离子抑制效应等原因．目前主要有 3 种方法处理数据缺失问题：a．不考虑具有缺失数据的肽段[55]；b．采用交叉搜索的策略[32, 56]；c．利用估计值填充缺失的数据[57−58]．显然，第一种处理方法简单，但不可避免地丢失了许多有用信息，并且可能导致蛋白质丰度比的过估计[57]，第二种方法可以有效地处理第一种数据缺失肽段，但是无法估计第二种数据缺失肽段的丰度表征，而第三种方法对两种数据缺失的肽段都适用，但是目前对这种方法的研究还不深入．

不管采用信号强度法还是图谱计数法定量，都存在共享肽段的丰度分配问题．目前也主要有 3 种处理方法：a．不考虑蛋白质中的共享肽段[59]；b. 在蛋白质群(peptide-sharing closure groups, PSCGs)中考虑共享肽段，并计算该群中所有蛋白质的总量[60]；c．使用合适的分配准则，把共享肽段的丰度分配到各个蛋白质中去[61−63]．显然，第一种方法丢失了许多共享肽段的定量数据，Zhang 等[63]评估了一系列的处理共享肽段的方法，结果表明丢弃共享肽段后的定量结果明显劣于考虑共享肽段后的结果．第二种方法揭示了 PSCGs 群中的蛋白质通常具有相似的生物学功能，但该方法只能够求出该群中所有蛋白质的丰度表征总值，且不适用于功能差异很大的蛋白质．第三种方法理论上最为合理，但选择合适的分配准则是关键．

2.6 统计学分析

统计学分析的主要目的是根据蛋白质的丰度比找出显著性差异表达的蛋白质．一般来说，估计的丰度差异表达不仅仅反映了生物样本中真实的差异，而且还包含了各种各样的随机误差，例如，生物重复样本的随机影响、仪器的测量误差等，所以需要利用统计假设检验来确定蛋白质是否存在显著性差异表达．

检验两组数据是否存在差异的一个成熟的统计学检验方法是 t 检验，但是 t 检验需要假设样本数据来自于一个正态分布，并且要求每种样本至少有 3 组重复实验[64−65]．实际上，无标记 LC-MS 实验测得的定量值一般不服从正态分布，但是对数变换可以使数据近似服从正态分布[52]．对于不服从正态分布的数据，也可以使用非参数假设检验方法，常用的非参数假设检验方法有置换检验和 K-S 检验，其优点是不用对数据的分布做任何假设[66]．若实验包括了两组以上的定量测量值，则可以使用方差分析和 Kruskal-Wallis 检验．

为了更好地适应无标记定量数据，学者们提出了很多新的统计学分析方法．其中，Tan 等[67]认为用多种统计学方法同时得到的显著差异表达的蛋白质更加可靠，所以利用 4 种不同的统计学方法检验同一批数据，得到了更可信的显著性差异表达的蛋白质．Pavelka 等[68]揭示了蛋白质组数据中的 NSAF 定量数据与微阵列数据的相似点，并把处理微阵列数据的 PLGEM(power law global error model)统计学方法[69]推广到 NASF 定量数据集上来，但这种方法需要至少 4 次重复质谱实验．针对 NASF 定量数据集，基于分层贝叶斯统计学方法，Choi 等[70]给出了一种新的统计学分析框架(QSpec)判别具有显著差异表达的蛋白质，这种方法对没有重复实验的数据仍然适用．

此外，多重检验问题也是蛋白质的差异分析需要考虑的问题．一般来说，对单个蛋白质的检验，P 值低于 0.05 就被认为具有显著性差异表达．当同时分析多个蛋白质时，便会出现问题．例如，对于包含 10 000 个蛋白质的定量数据集，设置 0.05 的 P 值阈值，理论上会产生 500 个假阳性结果．因此，为了限制假阳性结果的数量，需要进行校正多重检验．Bonferroni 方法通过控制总的Ⅰ型错误率(family-wise error rate)来校正多重检验，但是这种方法在降低假阳性率(false positive rate, FPR)的同时却大大增加了假阴性结果的数量[71]．另一种更加保守的方法是通过控制错误发现率(false discovery rate, FDR)来解决多重检验问题[12, 71−72]．尽管如此，定量蛋白质组数据的统计学分析问题还有待于进一步深入地研究[66, 73]．

经过统计学分析后，可以得到一系列显著性差异表达的蛋白质——称为候选生物标志物，真正的生物标志物可以通过多级反应检测技术(multiple reaction monitoring，MRM)[74−77]或基于抗体的检测技术对候选生物标志物验证，以及反复的临床验证得到．

3 无标记蛋白质组定量软件

表 1 列举了目前一些常用的无标记定量软件，软件对应的网络资源在表 2 中．这些软件各自有不同的特点和用途，大部分可以免费下载使用，部分还公开了源码．

Table 1 Summary of software tools for label-free proteomics quantification analysis

表 1 常用的无标记定量分析软件

策略	软件	操作系统	数据类型	数据格式	备注
流程一	SpecArray	Linux	FT-LTQ, OrbiTrap, Qtof	mzXML	整合在 TPP 中
	MsInspect	Linux,OSX,Windows	ESI-Tof, OrbiTrap, FT-LTQ, Qtof	mzXML	用户界面，命令行
	MapQuant	Linux,Windows	LCQ, FT-LTQ	mzXML, mzData, hmsXML	用户界面
	TOPP	Linux,OSX,Windows	LTQ, ESI-Tof	mzXML	用户界面
	PEPPeR	Linux,OSX,Windows	FT-LTQ, OrbiTrap	mzXML	整合在 Gene Pattern 中
	SuperHirn	Linux, OSX	FT-LTQ, OrbiTrap, Qtof	mzXML	整合在 TPP 中
	DeepQuanTR	Windows	LC-MALDI-MS	txt, mzXML, mzData	用户界面
	SIEVE	Windows	MS data from Thermo	raw	用户界面
	Expressionist	Windows	Thermo\ Bruker\ Warters 仪器	raw	用户界面
流程二	T3PQ	Windows	FT-LTQ, OrbiTrap	mzXML	命令行
	Census	Windows	FT-LTQ, LTQ, etc	mzXML, pepXML, MS	命令行
	IDEAL-Q	Windows	ESI-Tof, OrbiTrap, FT-LTQ, Qtof	mzXML	用户界面
	PeptideQuant	Windows	LC-ESI-MS	mzXML	Matlab 工具箱
	APEX	Windows	FT-LTQ, LTQ, Qtof	protXML	用户界面，图谱计数
	ProtQuant	Windows	FT-LTQ, LTQ, Qtof	sequantXML	用户界面，图谱计数

Table 2 Internet resources and references of label-free quantification tools

表 2 无标记定量软件对应的网络资源及相应的文献

软件	网址	文献	软件类型
SpecArray	http://tools.proteomecenter.org/software.php	[12]	公开源码, C
MsInspect	http://proteomics.fhcrc.org/CPL/home.html	[7]	公开源码, Java
MapQuant	http://arep.med.harvard.edu/MapQuant/	[78]	免费使用, C++
TOPP	http://open-ms.sourceforge.net	[79]	公开源码, C++
PEPPeR	http://www.broad.mit.edu/cancer/software/genepattern/	[80]	公开源码, Perl 和 R
SuperHirn	http://tools.proteomecenter.org/software.php	[81]	公开源码, C++
DeepQuanTR	http://www.pharma.ethz.ch/institute_groups/institute_groups/biomacromolecules/deepquantr	[28]	免费使用, VB
SIEVE	http://www.thermo.com/	-	商业软件
Expressionist	http://www.genedata.com/	-	商业软件
T3PQ	http://fqms.svn.sourceforge.net/svnroot/fqms	[44]	公开源码, Python
Census	http://fields.scripps.edu/census/index.php	[82]	免费使用, Java
IDEAL-Q	http://ms.iis.sinica.edu.tw/IDEAL-Q/	[32]	免费使用, .NET
PeptideQuant	http://bioinformatics.ust.hk/PeptideQuant/peptidequant.htm	[33]	公开源码, Matlab
APEX	http://pfgrc.jcvi.org/index.php/bioinformatics/apex.html	[83]	免费使用, Java
ProtQuant	http://www.agbase.msstate.edu/tools.html	[84]	免费使用, Java

对于流程一，MS 肽段特征的信息提取、不同 LC-MS 间肽段特征的对齐和肽段 / 蛋白质序列匹配是软件需考虑的问题．在特征信息提取方面，大多数软件都用到了肽段的天然同位素分布信息，但是 DeepQuanTR 除外，由于 DeepQuanTR 只处理 MALDI-TOF 类的质谱数据，而 MALDI 离子源主要产生单电荷的肽段离子，所以其肽段特征提取算法相对比较简单．在肽段特征对齐方面，色谱保留时间对齐是关键．SpecArray、MsInspect 和 SuperHirn 都是利用拟合、回归的方法，通过估计两两 LC-MS 实验间的保留时间校正曲线实现 RT 对齐；DeepQuanTR 则是利用 R-Square 值来衡量两

次实验中的肽段特征匹配的相似性，通过求取最高相似性的特征匹配来实现；MapQuant 没有考虑 RT 对齐，无法识别不同实验间的相同肽段特征．在肽段 / 蛋白质序列匹配方面，MsInspect 利用所有实验的鉴定结果，构建精确质量时间(AMT)标签数据库，通过 AMT 标签数据库搜索实现肽段特征的肽段 / 蛋白质序列匹配，PEPPeR 的肽段特征对齐和序列匹配算法是同时进行的，该软件把所有 LC-MS 实验的鉴定肽段看作一个整体，通过质荷比误差容限、匹配打分得到了一小部分肽段特征的序列匹配，在此基础上，利用实验间的质荷比误差校正和 RT 校正、高斯混合模型聚类算法实现了肽段特征的对齐，MapQuant、TOPP 和 SuperHirn 则通过设置质荷比窗口和保留时间窗口，实现了肽段特征与本次 LC-MS 实验的鉴定肽段的匹配，但能够匹配到序列的肽段特征有限，而 SpecArray 没有考虑序列匹配．就软件本身特点而言，TOPP 对不同的数据处理步骤都实现了多种算法，例如，图谱去噪算法包含了小波变换去噪、高斯去噪、Savitzky-Golay 去噪等，每个定量步骤用一个模块来实现，用户可以根据数据特点和自身需求定制出不同的数据分析流程，另外，该软件公开了 C++源码，用户可以在此平台上改进某些算法．

流程二主要包括信号强度法和图谱计数法两种定量方法．一般来说，信号强度法定量是利用某个实验的鉴定结果，返回到这个实验的一级图谱中提取肽段的 XIC，进而完成肽段和蛋白质的定量，如 T3PQ、Census 和 PeptideQuant．这种定量策略的不足之处是最终只能够比较分析不同实验间共同鉴定到的肽段和蛋白质．为了解决这个问题，IDEAL-Q 采用了交叉搜索策略以及有效的图谱数据验证准则，充分利用了所有实验的鉴定肽段，可以得到更多的肽段定量结果．PeptideQuant 只是肽段水平的定量软件，但是其定量方法很好地解决了肽段质谱峰和 XIC 峰重叠的问题．APEX 和 ProtQuant 是基于图谱计数法的定量软件．其中，APEX 实现了蛋白质检测效率预测的机器学习方法，把蛋白质效率预测校正的图谱数作为蛋白质的丰度表征．ProtQuant 使用肽段的 Sequest 鉴定分值(XCorr)之和来校正对应蛋白质的图谱数，并采用了一种新的处理数据缺失问题的算法．

以上两种数据处理流程都有各自相同或不同的数据处理步骤，而不同的处理步骤中又有多种算法可以选择，如何选取其中几种性能较优的算法组合设计新的定量软件是今后的一项重要工作．

4 总结与展望

本文介绍了蛋白质质谱分析的无标记定量问题，总结了无需 / 需要鉴定结果两种定量方法及其对应的数据处理流程，分析比较了这两种定量方法的异同及优缺点，详细讨论了数据处理流程涉及的主要算法，列举了一些常用的无标记定量软件及其对应的网络资源．

蛋白质质谱分析的无标记定量已成为定量蛋白质组学中常用的技术之一，但与有标记定量技术相比，仍存在诸如可重复性差、定量准确性低等问题．从长远来看，实验技术和实验设备的不断发展无疑会有助于无标记定量性能的提高；但就近期来讲，改进定量算法不失为一个很好的选择．本文认为，应该从以下几个方面更深入地研究无标记定量算法：

a．算法评估．数据处理流程中的每个步骤均有很多可选的算法，算法的不同组合可以得到不同的定量方法，但是目前缺少对这些定量方法的系统评估，尤其是在高精度质谱数据中的评估．通过评估可以给出一组或几组针对不同质谱数据的最优方法．

b．改进无需鉴定结果的定量方法中各步骤的算法．对于谱峰检测，现有算法大都是针对低精度的质谱数据开发的，如何从高精度的质谱数据中高效地检测出肽段峰还很少讨论；对于保留时间对齐，利用原始谱数据实现色谱保留时间对齐可能是今后的研究热点．

c．整合信号强度法定量和图谱计数法定量的定量结果．图谱计数法具有运算速度快、在一定丰度范围内定量结果灵敏度高等特点，而信号强度法在处理鉴定图谱数很少的蛋白质时具有图谱计数无法比拟的优势．整合这两种方法的结果可以提高定量结果的可靠性．

d．进一步发展肽段 / 蛋白质定量指标．定量信息提取直接影响了定量算法的准确性，不断挖掘肽段丰度与图谱信息之间更深层次的关系、发展新的肽段 / 蛋白质定量指标是提升定量算法的关键．此外，对于低丰度肽段 / 蛋白质的定量信息提取也是需要着重研究的问题．

e．考虑蛋白质丰度比推断中的肽段数据缺失问题和共享肽段问题．这两个问题都会导致蛋白质丰度比的低估或高估．

f. 发展适合蛋白质定量数据特点的统计学方法．由于实验成本、蛋白质样本获取等的限制，可得到的重复实验的数据很少，这对蛋白质丰度的差异显著性分析带来很大的挑战．另外，对多重检验带来的假阳性结果的控制也需要进一步研究．

参 考 文 献

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

[2] Elliott M H, Smith D S, Parker C E, *et al*. Current trends in quantitative proteomics. J Mass Spectrom, 2009, **44**(12):1637−1660

[3] Zhu W, Smith J W, Huang C M. Mass spectrometry-based label-free quantitative proteomics. J Biomed Biotechnol, 2010, **2010**:840518

[4] Schmidt A, Gehlenborg N, Bodenmiller B, *et al*. An integrated, directed mass spectrometric approach for in-depth characterization of complex peptide mixtures. Mol Cell Proteomics, 2008, **7**(11): 2138−2150

[5] May D, Fitzgibbon M, Liu Y, *et al*. A platform for accurate mass and time analyses of mass spectrometry data. J Proteome Res, 2007, **6**(7): 2685−2694

[6] America A H, Cordewener J H. Comparative LC-MS: a landscape of peaks and valleys. Proteomics, 2008, **8**(4): 731−749

[7] 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

[8] Geurts P, Fillet M, de Seny D, *et al*. Proteomic mass spectra classification using decision tree based ensemble methods. Bioinformatics, 2005, **21**(14): 3138−3145

[9] Liu Q, Sung A H, Qiao M, *et al*. Comparison of feature selection and classification for MALDI-MS data. BMC Genomics, 2009, **10**(Suppl 1): S3

[10] Cannataro M, Cuda G, Gaspari M, *et al*. The EIPeptiDi tool: enhancing peptide discovery in ICAT-based LC MS/MS experiments. BMC Bioinformatics, 2007, **8**: 255

[11] Tabb D L, Vega-Montoto L, Rudnick P A, *et al*. Repeatability and reproducibility in proteomic identifications by liquid chromatography-tandem mass spectrometry. J Proteome Res, 2010, **9**(2): 761−776

[12] 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

[13] Jimmy K. Eng A L M, Yates R 3rd. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. J Am Soc Mass Spectrometry, 1994, **5**(11): 14

[14] 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

[15] Elias J E, Gygi S P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. Nat Methods, 2007, **4**(3): 207−214

[16] Keller A, Nesvizhskii A I, Kolker E, *et al*. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. Anal Chem, 2002, **74**(20): 5383−5392

[17] Renard B Y, Timm W, Kirchner M, *et al*. Estimating the confidence of peptide identifications without decoy databases. Anal Chem, 2010, **82**(11): 4314−4318

[18] Tabb D L, McDonald W H, Yates J R, 3rd. DTASelect and Contrast: tools for assembling and comparing protein identifications from shotgun proteomics. J Proteome Res, 2002, **1**(1): 21−26

[19] Yasui Y, Pepe M, Thompson M L, *et al*. A data-analytic strategy for protein biomarker discovery: profiling of high-dimensional proteomic data for cancer detection. Biostatistics, 2003, **4**(3):449−463

[20] Coombes K R, Tsavachidis S, Morris J S, *et al*. Improved peak detection and quantification of mass spectrometry data acquired from surface-enhanced laser desorption and ionization by denoising spectra with the undecimated discrete wavelet transform. Proteomics, 2005, **5**(16): 4107−4117

[21] Du P, Kibbe W A, Lin S M. Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching. Bioinformatics, 2006, **22**(17): 2059−2065

[22] Mantini D, Petrucci F, Pieragostino D, *et al*. LIMPIC: a computational method for the separation of protein MALDI-TOF-MS signals from noise. BMC Bioinformatics, 2007, **8**: 101

[23] Cruz-Marcelo A, Guerra R, Vannucci M, *et al*. Comparison of algorithms for pre-processing of SELDI-TOF mass spectrometry data. Bioinformatics, 2008, **24**(19): 2129−2136

[24] Yang C, He Z, Yu W. Comparison of public peak detection algorithms for MALDI mass spectrometry data analysis. BMC Bioinformatics, 2009, **10**: 4

[25] Wu L C, Chen H H, Horng J T, *et al*. A novel preprocessing method using Hilbert Huang Transform for MALDI-TOF and SELDI-TOF mass spectrometry data. PLoS One, 2010, **5**(8): e12493

[26] McLerran D F, Feng Z, Semmes O J, *et al*. Signal detection in high-resolution mass spectrometry data. J Proteome Res, 2008, **7**(1): 276−285

[27] Zhang S, DeGraba T J, Wang H, *et al*. A novel peak detection approach with chemical noise removal using short-time FFT for prOTOF MS data. Proteomics, 2009, **9**(15): 3833−3842

[28] Fugmann T, Neri D, Roesli C. DeepQuanTR: MALDI-MS-based label-free quantification of proteins in complex biological samples. Proteomics, 2010, **10**(14): 2631−2643

[29] Jiang W, Qiu Y, Ni Y, *et al*. An automated data analysis pipeline for GC-TOF-MS metabonomics studies. J Proteome Res, 2010, **9**(11): 5974−5981

[30] Chelius D, Bondarenko P V. Quantitative profiling of proteins in complex mixtures using liquid chromatography and mass spectrometry. J Proteome Res, 2002, **1**(4): 317−323

[31] Higgs R E, Knierman M D, Gelfanova V, *et al*. Comprehensive label-free method for the relative quantification of proteins from

biological samples. J Proteome Res, 2005, **4**(4): 1442−1450

[32] Tsou C C, Tsai C F, Tsui Y H, *et al*. IDEAL-Q, an automated tool for label-free quantitation analysis using an efficient peptide alignment approach and spectral data validation. Mol Cell Proteomics, 2010, **9**(1): 131−144

[33] Yang C, Yu W. A regularized method for peptide quantification. J Proteome Res, 2010, **9**(5): 2705−2712

[34] Old W M, Meyer-Arendt K, Aveline-Wolf L, *et al*. Comparison of label-free methods for quantifying human proteins by shotgun proteomics. Mol Cell Proteomics, 2005, **4**(10): 1487−1502

[35] Liu H, Sadygov R G, Yates J R, 3rd. A model for random sampling and estimation of relative protein abundance in shotgun proteomics. Anal Chem, 2004, **76**(14): 4193−4201

[36] Ishihama Y, Oda Y, Tabata T, *et al*. Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein. Mol Cell Proteomics, 2005, **4**(9): 1265−1272

[37] Zybailov B, Mosley A L, Sardiu M E, *et al*. Statistical analysis of membrane proteome expression changes in Saccharomyces cerevisiae. J Proteome Res, 2006, **5**(9): 2339−2347

[38] Lu P, Vogel C, Wang R, *et al*. Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. Nat Biotechnol, 2007, **25**(1): 117−124

[39] Sun A, Zhang J, Wang C, *et al*. Modified spectral count index (mSCI) for estimation of protein abundance by protein relative identification possibility (RIPpro): a new proteomic technological parameter. J Proteome Res, 2009, **8**(11): 4934−4942

[40] Griffin N M, Yu J, Long F, *et al*. Label-free, normalized quantification of complex mass spectrometry data for proteomic analysis. Nat Biotechnol, 2010, **28**(1): 83−89

[41] Zybailov B, Coleman M K, Florens L. Correlation of relative abundance ratios derived from peptide ion chromatograms and spectrum counting for quantitative proteomic analysis using stable isotope labeling. Anal Chem, 2005, **77**(19): 6218−6224

[42] Asara J M, Christofk H R, Freimark L M, *et al*. A label-free quantification method by MS/MS TIC compared to SILAC and spectral counting in a proteomics screen. Proteomics, 2008, **8**(5): 994−999

[43] Xia Q, Wang T, Park Y, *et al*. Differential quantitative proteomics of Porphyromonas gingivalis by linear ion trap mass spectrometry: non-label methods comparison, q-values and LOWESS curve fitting. Int J Mass Spectrom, 2007, **259**(1−3): 105−116

[44] Grossmann J, Roschitzki B, Panse C, *et al*. Implementation and evaluation of relative and absolute quantification in shotgun proteomics with label-free methods. J Proteomics, 2010, **73**(9): 1740−1746

[45] Vandenbogaert M, Li-Thiao-Te S, Kaltenbach H M, *et al*. Alignment of LC-MS images, with applications to biomarker discovery and protein identification. Proteomics, 2008, **8**(4):650−672

[46] Podwojski K, Fritsch A, Chamrad D C, *et al*. Retention time alignment algorithms for LC/MS data must consider non-linear shifts. Bioinformatics, 2009, **25**(6): 758−764

[47] Marc Kirchner B S, Hanno Steen, Judith A J. Steen. amsrpm: robust point matching for retention time aligment of LC/MS data with R. J Statistical Software, 2007, **18**(4): 1−12

[48] Eilers P H. Parametric time warping. Anal Chem, 2004, **76**(2): 404−411

[49] Christin C, Hoefsloot H C, Smilde A K, *et al*. Time alignment algorithms based on selected mass traces for complex LC-MS data. J Proteome Res, 2010, **9**(3): 1483−1495

[50] Kong X, Reilly C. A Bayesian approach to the alignment of mass spectra. Bioinformatics, 2009, **25**(24): 3213−3220

[51] Bolstad B M, Irizarry R A, Astrand M, *et al*. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics, 2003, **19**(2): 185−193

[52] Callister S J, Barry R C, Adkins J N, *et al*. Normalization approaches for removing systematic biases associated with mass spectrometry and label-free proteomics. J Proteome Res, 2006, **5**(2): 277−286

[53] Kultima K, Nilsson A, Scholz B, *et al*. Development and evaluation of normalization methods for label-free relative quantification of endogenous peptides. Mol Cell Proteomics, 2009, **8**(10): 2285−2295

[54] Carrillo B, Yanofsky C, Laboissiere S, *et al*. Methods for combining peptide intensities to estimate relative protein abundance. Bioinformatics, 2010, **26**(1): 98−103

[55] Oberg A L, Mahoney D W, Eckel-Passow J E, *et al*. Statistical analysis of relative labeled mass spectrometry data from complex samples using ANOVA. J Proteome Res, 2008, **7**(1): 225−233

[56] Andreev V P, Li L, Cao L, *et al*. A new algorithm using cross-assignment for label-free quantitation with LC-LTQ-FT MS. J Proteome Res, 2007, **6**(6): 2186−2194

[57] Wang P, Tang H, Zhang H, *et al*. Normalization regarding non-random missing values in highthroughput mass spectrometry data. Pac Symp Biocomput, 2006: 315−326

[58] Karpievitch Y, Stanley J, Taverner T, *et al*. A statistical framework for protein quantitation in bottom-up MS-based proteomics. Bioinformatics, 2009, **25**(16): 2028−2034

[59] Usaite R, Wohlschlegel J, Venable J D, *et al*. Characterization of global yeast quantitative proteome data generated from the wild-type and glucose repression saccharomyces cerevisiae strains: the comparison of two quantitative methods. J Proteome Res, 2008, **7**(1): 266−275

[60] Jin S, Daly D S, Springer D L, *et al*. The effects of shared peptides on protein quantitation in label-free proteomics by LC/MS/MS. J Proteome Res, 2008, **7**(1): 164−169

[61] Zybailov B, Rutschow H, Friso G, *et al*. Sorting signals, N-terminal modifications and abundance of the chloroplast proteome. PLoS One, 2008, **3**(4): e1994

[62] Liu W L, Coleman R A, Grob P, *et al*. Structural changes in TAF4b-TFIID correlate with promoter selectivity. Mol Cell, 2008, **29**(1): 81−91

[63] Zhang Y, Wen Z, Washburn M P, *et al*. Refinements to label free

proteome quantitation: how to deal with peptides shared by multiple proteins. Anal Chem, 2010, **82**(6): 2272−2281

[64] Bantscheff M, Schirle M, Sweetman G, *et al*. Quantitative mass spectrometry in proteomics: a critical review. Anal Bioanal Chem, 2007, **389**(4): 1017−1031

[65] Zhang B, VerBerkmoes N C, Langston M A, *et al*. Detecting differential and correlated protein expression in label-free shotgun proteomics. J Proteome Res, 2006, **5**(11): 2909−2918

[66] Listgarten J, Emili A. Statistical and computational methods for comparative proteomic profiling using liquid chromatography-tandem mass spectrometry. Mol Cell Proteomics, 2005, **4**(4):419−434

[67] Tan N C, Fisher W G, Rosenblatt K P, *et al*. Application of multiple statistical tests to enhance mass spectrometry-based biomarker discovery. BMC Bioinformatics, 2009, **10**: 144

[68] Pavelka N, Fournier M L, Swanson S K, *et al*. Statistical similarities between transcriptomics and quantitative shotgun proteomics data. Mol Cell Proteomics, 2008, **7**(4): 631−644

[69] Pavelka N, Pelizzola M, Vizzardelli C, *et al*. A power law global error model for the identification of differentially expressed genes in microarray data. BMC Bioinformatics, 2004, **5**: 203

[70] Choi H, Fermin D, Nesvizhskii A I. Significance analysis of spectral count data in label-free shotgun proteomics. Mol Cell Proteomics, 2008, **7**(12): 2373−2385

[71] Gutstein H B, Morris J S, Annangudi S P, *et al*. Microproteomics: analysis of protein diversity in small samples. Mass Spectrom Rev, 2008, **27**(4): 316−330

[72] Li Q, Roxas B A. An assessment of false discovery rates and statistical significance in label-free quantitative proteomics with combined filters. BMC Bioinformatics, 2009, **10**: 43

[73] Mueller L N, Brusniak M Y, Mani D R, *et al*. An assessment of software solutions for the analysis of mass spectrometry based quantitative proteomics data. J Proteome Res, 2008, **7**(1): 51−61

[74] Lange V, Picotti P, Domon B, *et al*. Selected reaction monitoring for quantitative proteomics: a tutorial. Mol Syst Biol, 2008, **4**(10): 222

[75] Schiess R, Wollscheid B, Aebersold R. Targeted proteomic strategy for clinical biomarker discovery. Mol Oncol, 2009, **3**(1): 33−44

[76] Oh J H, Pan S, Zhang J, *et al*. MSQ: a tool for quantification of proteomics data generated by a liquid chromatography/matrix-assisted laser desorption/ionization time-of-flight tandem mass spectrometry based targeted quantitative proteomics platform. Rapid Commun Mass Spectrom, 2010, **24**(4): 403−408

[77] Lange V, Malmstrom J A, Didion J, *et al*. Targeted quantitative analysis of Streptococcus pyogenes virulence factors by multiple reaction monitoring. Mol Cell Proteomics, 2008, **7**(8): 1489−1500

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

[79] Kohlbacher O, Reinert K, Gropl C, *et al*. TOPP--the OpenMS proteomics pipeline. Bioinformatics, 2007, **23**(2): e191−197

[80] Jaffe J D, Mani D R, Leptos K C, *et al*. PEPPeR, a platform for experimental proteomic pattern recognition. Mol Cell Proteomics, 2006, **5**(10): 1927−1941

[81] 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

[82] Park S K, Venable J D, Xu T, *et al*. A quantitative analysis software tool for mass spectrometry-based proteomics. Nat Methods, 2008, **5**(4): 319−322

[83] Braisted J C, Kuntumalla S, Vogel C, *et al*. The APEX Quantitative Proteomics Tool: generating protein quantitation estimates from LC-MS/MS proteomics results. BMC Bioinformatics, 2008, **9**: 529

[84] Bridges S M, Magee G B, Wang N, *et al*. ProtQuant: a tool for the label-free quantification of MudPIT proteomics data. BMC Bioinformatics, 2007, **8**(Suppl 7): S24

Development of Algorithms for Mass Spectrometry-based Label-free Quantitative Proteomics*

ZHANG Wei[1)], ZHANG Ji-Yang[1)], LIU Hui[1)], SUN Han-Chang[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 As a main approach for disease related biomarkers discovery, quantitative research has become a hot topic in proteomics. With the development of experimental methods, the quantitative data processing methods are updated and improved constantly. Two categories of label-free quantitative algorithms are introduced at first, including database-free methods and database searching-based methods. Then, the analysis strategies and the details of the two kinds of methods are described, the advantages and disadvantages are also investigated, the frequently used tools and the corresponding internet resources are summarized. At last, some suggestions for improving the data processing of label-free quantitative proteomics are proposed.

Key words quantitative proteomics, mass spectrometry, label-free quantification, quantitative algorithms, statistic analysis

DOI: 10.3724/SP.J.1206.2010.00560

* This work was supported by grants from The National Natural Science Fundation of China (31000587) and State Key Laboratory of Proteomics (SKLP-O201004).

**Corresponding author.

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

Received: November 1, 2010 Accepted: January 6, 2011