PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2013, 40(3): 281~292
www.pibb.ac.cn

定量蛋白质组学无标记定量方法的研究进展 *

武 鹏 贺福初 姜 颖 **
(蛋白质组学国家重点实验室，北京蛋白质组研究中心，军事医学科学院放射与辐射医学研究所，北京 102206)

摘要 依靠质谱技术的蛋白质组学快速发展，寻求速度快、重复性好以及准确度高的定量方法是该领域的一项艰巨任务，定量蛋白质组学分支领域应运而生．其中，无标记定量方法以其样品制备简单、耗材费用低廉以及结果数据分析便捷等优点渐露锋芒．无标记定量方法通常分为信号强度法和谱图计数法两大类．本文在这两种无标记定量方法计算原理的基础上，针对各种常用的无标记定量方法及最新进展做一个较为全面的介绍，并将详细讨论两类方法的异同点，以及目前蛋白质组学中无标记定量方法所面临的主要挑战，希望能为这一领域的研究人员在选择无标记定量方法时提供一个合理的参考．

关键词 定量蛋白质组学，生物信息学，无标记定量，信号强度，谱图计数，共享肽段
学科分类号 Q51，Q811.4 **DOI**: 10.3724/SP.J.1206.2012.00055

蛋白质组学旨在鉴定出一个细胞、组织或器官中全部的蛋白质信息，质谱作为蛋白质组学研究的支撑技术发挥着核心作用[1]．如何从大规模质谱数据中提取蛋白质表达水平即蛋白质定量信息一直是蛋白质组学关注的热点．在定量蛋白质组学中，基于标记的定量方法具有较高的准确性，已经得到广泛应用，其中包括标记肽段的同位素标签法(isobaric tags for relative and absolute quantitation，iTRAQ)[2]、标记蛋白质的同位素亲和标签法(isotope- coded affinity tags，ICAT)[3]以及在细胞培养过程中引入稳定同位素标记法(stable isotope labels with amino acids in cell culture，SILAC)[4]等．然而，基于标记的定量方法存在一些缺陷，如样品制备复杂、同位素标记试剂昂贵、数据分析软件要求特殊等．近年来，研究人员发现蛋白质丰度与质谱数据一级谱图肽段峰强度或二级谱图数目呈现一定相关性，使无标记定量成为了定量蛋白质组学的另一有效方法．无标记定量在一定程度上可以克服标记定量方法的缺点，在单次实验中可定量蛋白质数据更多．同时也有研究表明，无标记定量方法与基于标记的定量方法可达到相似的准确度[5]．因此，无标记定量方法已经逐渐地应用于蛋白质表达谱的建立以及生物标志物的发现等研究中．

基于质谱数据的无标记定量方法通常分成两类．其一，基于一级谱图母离子强度，如谱图峰高、谱图峰面积和谱图峰容量等，称为信号强度法，又叫做离子强度法或曲线面积法；其二，基于肽段匹配的二级谱图数目，称为谱图计数法．两类无标记定量方法计算流程如图 1 所示．本文将详细介绍这两类常用的无标记定量方法的原理和最新进展，并比较信号强度法和谱图计数法的异同点和优缺点，最后阐述无标记定量方法面临的两大挑战，即共享肽段处理和统计检验．

* 国家重点基础研究发展计划(973)(2010CB912700) 资助项目．
** 通讯联系人．
Tel: 010-80727777-1213, E-mail: jiangying304@hotmail.com
收稿日期：2012-04-17，接受日期：2012-06-01

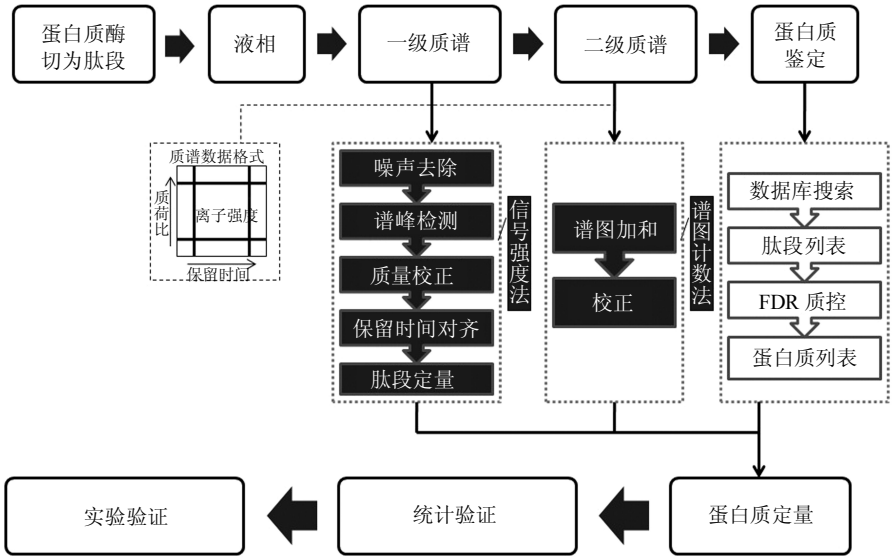

Fig. 1　General workflow of label-free quantitative methods

图 1　无标记定量方法的主要流程

1　两类常见的无标记定量方法

1.1　信号强度法

经典的蛋白质组学流程首先用胰酶将蛋白质酶切成肽段，然后经过液相色谱分离，根据肽段的疏水性、离子强度、等电点的差异进行分离洗脱，接着利用电喷雾离子源产生气相离子．质谱仪根据它们的质荷比将其分离并记录，这一阶段称为母离子扫描．然后按照不同质谱仪设定的规则，通常选择丰度最高的 5～9 个母离子依次进行二级碎裂，得到二级碎片离子谱图，再返回至母离子扫描，进入下一循环．这样的一种质谱数据获取方式被称为数据依赖获取(data dependent acquisition, DDA)模式[1,6]，也是最为常见的一种模式．在一级质谱图中，每一个母离子都包括三维信息，即液相色谱保留时间、质荷比以及离子强度(图 1)．由于电喷雾离子化得到的母离子信号强度与离子浓度存在相关性[7]，每个母离子都是离子化的肽段，因此在一级谱图中提取出鉴定肽段对应的离子峰强度即可以反映出肽段的丰度，常用来表示离子峰强度的参数有峰高度、峰面积和峰容积等．在提取离子信号强度之前，往往需要对原始质谱数据进行预处理，通常这也是信号强度法最重要且最复杂的步骤．

1.1.1　一级谱图预处理．

在质谱母离子扫描阶段，某一个肽段会在一个时间段内从液相柱洗脱下来，导致其产生的信号往往跨越一定的液相色谱保留时间范围．其次，由于肽段的不同同位素构成以及离子化所带的不同电荷数等因素，肽段产生的信号可能跨越一定的质荷比范围．因此，一级谱图不同区域的信号可能来自相同肽段．在单次质谱实验中，肽段数量高达数百时，保留时间的偏差可能达到数分钟．同时质谱仪器的系统噪声也会引起保留时间、质荷比以及信号强度的扰动变化，另外一些化学噪声也会产生干扰，这些因素都会给提取肽段信号强度带来困难．因此，在进行无标记定量计算之前，需要对原始数据进行预处理．

一级谱图数据预处理通常包括三个步骤：首先从原始谱图数据中去除噪声和基线，然后进行谱峰检测，最后进行质量校正和保留时间对齐．目前，预处理的每一个步骤，都已有多种算法进行解决．张伟等[8]对该类算法的研究进行了较为全面的综述，在此不再赘述．然而，在该类算法研究中，不同算法处理不同性质质谱数据的性能差异，相邻步

骤之间的算法衔接，每一步算法的处理带给最终结果的错误率等问题至今仍不清楚，需要深入研究.

针对一级谱图信号强度的定量方法，本文将选取一些代表性的软件工具加以说明(表 1).

Table 1 Tools for label-free quantification based on signal intensity

表 1 基于信号强度的无标记定量工具

工具	特点	质谱类型	优势	网址	文献
T3PQ	3 个最高峰强度的平均	FT-LTQ, Orbitrap	开源软件，使用简便	http://fqms.svn.sourceforge.net/svnroot/fqms	[9]
Quoil	Biweight 算法应用	FT-LTQ, Orbitrap	数据处理速度快	------	[10]
msInspect	数据二维特性应用	TOF, FT-LTQ 等	开源软件，用户可扩展算法	http://proteomics.fhcrc.org/CPL/home.html	[11]
Superhirn	谱图相似性分析	FT-LTQ, Orbitrap	不同实验室数据可整合	http://tools.proteomecenter.org/software.php	[12]
Census	肽段加权测量、动态峰搜索	LTQ, FT-LTQ 等	应用范围广，支持高、低精度质谱数据	http://fields.scripps.edu/census/index.php	[13]
IDEAL-Q	交叉搜索、SCI 过滤标准	TOF, FT-LTQ, Orbitrap	增加可定量的肽段数目	http://ms.iis.sinica.edu.tw/IDEAL-Q	[14]
PeptideQuant	同位素峰分布、洗脱峰平滑去噪	FT-LTQ, Orbitrap	同时考虑计算偏性和变化	http://bioinformatics.ust.hk/PeptideQuant	[15]

1.1.2 常见的信号强度法.

最早，Chelius 等[16]提出肽段信号峰面积与肽段丰度呈正比并进行了实验验证. 实验人员将 10^{-14}～10^{-12} mol 范围的肌红蛋白酶切为肽段后经 LC-MS/MS 分析，利用一级谱图的肽段峰面积进行无标记定量计算，定量结果与原始蛋白质浓度呈线性相关(r^2=0.991). 然后将已知浓度的马肌红蛋白加入人血清样品，无标记定量结果与实际浓度相差低于 16%，进一步验证了该方法的可行性. 此后，一系列的基于一级肽段信号强度的无标记定量方法陆续出现.

Silva 等[17]经研究发现蛋白质浓度与其对应肽段的 3 个最高离子峰强度的平均值相关性最好. 此外，结合内标使用，此方法计算的 6 种不同浓度蛋白质的绝对丰度错误率在±15%以内，与已知蛋白质浓度相关性较好(r^2=0.9939). 值得注意的是，该研究是在质谱仪 LC-MS^E 模式下进行的. 但是，Grossmann 等[9]利用相似原理开发了 T3PQ(Top 3 Protein Quantification)无标记定量工具并成功地应用于 DDA 模式下的相对和绝对定量计算.

程序 Quoil(quantification without isotope labeling)能够快速处理大量的质谱原始文件[10]. 对鉴定到的肽段，计算不同样品间谱峰面积的比值，然后利用相应的肽段比值推断蛋白质比值. Quoil 利用 biweight 算法代替简单的数学平均值计算，使某些极端值的影响降到最小. 同时在后续统计分析中，运用 step-down 算法调整统计显著性的概率阈值，从而降低了假阳性率. 最后，研究人员运用 Quoil 在不同的复杂样品蛋白质组分析中，展现了良好的重复性和线性.

Bellew 等[11]针对高分辨率仪器开发了软件 msInspect，利用一系列算法进行基于一级质谱信号强度的无标记定量. msInspect 的信号处理模块利用数据的二维特性鉴定共洗脱同位素峰并基于同位素峰分布的相似性对肽段分组；校正方法则利用了不同实验批次之间色谱保留时间非线性关系；归一化过程借鉴基因组学芯片分析的方法以消除不同批次一级信号强度的系统偏差. msInspect 提供一个开源平台，使用者可以在软件中添加其他合适算法和流程. 质谱数据精确质量和时间分析平台 msInspect/AMT 结合了 LC-MS/MS 的鉴定结果和高分辨率质谱精确质量和保留时间定位，集合了基于 LC-MS/MS 的 CPAS (Computational Proteomics Analysis System)[18]和 msInspect 数据分析平台.

Superhirn 是一款专门针对高精度质谱仪数据的无标记定量软件[12]. 该工具囊括多个模块用于处理一级高精度质谱数据，包括特征提取及定量、基于谱图相似性的色谱保留时间对齐、多数据集一级谱图校正和强度归一化等. 其中谱图的相似性分析是该工具的一大亮点，可将不同批次的一级谱图数

据整合至一张 MasterMap 中，作为后续数据分析的基础．不同实验 MasterMap 的整合也为一级质谱数据整合以及一级质谱特性比较提供了可能．

John Yates 实验室开发的定量软件 Census[13]，其在标记和无标记策略、一级质谱与二级质谱扫描以及高低分辨率质谱数据中均可应用．Census 利用多种算法如加权肽段测量、动态峰搜寻以及统计过滤等，解决了低质量谱图计算障碍并提高了定量效率．在无标记定量中，Census 利用色谱校正进行定量分析，成功地用于了具有 10 个标准蛋白质样品的 4 次技术重复质谱分析中．

IDEAL-Q 定量工具[14]充分利用了所有实验批次鉴定到的肽段，运用线性回归方程和碎片回归优化方程的算法预测出肽段的洗脱时间，从而提取出该肽段的离子流色谱峰，实现了并行批次交叉搜索的目的．而且检测到的肽段峰需进一步通过 SCI (signal-to-noise ratio， charge state， and isotopic distribution)标准过滤掉噪音数据．在大肠杆菌细胞裂解液样品差异蛋白质组学分析中，IDEAL-Q 计算结果与期望值相关系数达 0.996．同时 IDEAL-Q 还可以兼容不同样品预分离策略，支持多种归一化方案．

Yang 等[15]开发了 PeptideQuant，利用肽段同位素峰分布和肽段洗脱谱平滑去噪，有效地解决了肽段重叠和峰强容易变化的两大难题．利用肽段同位素峰分布信息对肽段信号进行了优化，与其他方法相比，对不同样品中肽段丰度的计算更加准确．同时该方法还设置了与峰强变化相关的参数，折中考虑计算的偏性和变化．

总之，随着质谱仪器的发展，无标记定量数据处理流程也在不断更新进步．目前生物信息领域越来越重视数据处理算法的研究，对数据质量控制和参数优化提供支持．例如，程序 Quoil 灵活地应用了 biweight 和 step-down 算法，使无标记定量结果的线性和重复性得到提高．对信号强度法的每个关键步骤合理选用算法均可使定量准确性和灵敏度得到改善．同时，在程序设计时融入并行计算技术可大大加快程序运行速度，对处理大规模数据优势更加明显，是今后改善无标记定量方法的一个重要方向．

1.2 谱图计数法

通常，样品中某蛋白质丰度越高，则酶切产生的肽段也会越多．酶切肽段越多，经质谱分析，鉴定到的唯一匹配肽段(该肽段仅匹配至一个蛋白质)会越多，其对应的二级谱图数也会越多[19]．这即是基于谱图计数无标记定量方法的基础，因此人们尝试利用谱图计数作为一种简单的方法用于蛋白质的相对定量．谱图计数(spectral counts，SpC)模型直接统计蛋白质所有肽段的二级谱图数总和来反映蛋白质丰度．Ryu 等[20]在分别研究序列覆盖度、肽段数和谱图数与蛋白质丰度相关性分析后发现，仅谱图数与相对蛋白质丰度呈强线性相关．然而通过对谱图计数进行不同的校正，可以有效地改善此类方法的性能．表 2 列出了一些常用的谱图计数方法公式．

Table 2 The functions of common quantitative methods based spectral-counting

表 2 常用谱图计数方法的计算公式

方法	公式	注释	优点	文献
SpCn	$SpCn_i=\frac{SpC_i}{\Sigma SpC}$	SpC 为蛋白质的谱图数目.	消除总谱图数差异	[21]
NSAF	$NSAF_i=\frac{SpC_i/L_i}{\Sigma(SC/L)}$	L 为蛋白质的序列长度.	考虑蛋白质长度	[22-23]
APEX	$APEX_i=\frac{SpC_i\times P_i}{O_i\times[\Sigma(SpC\times P/O)]}\times C$	P 为蛋白质的鉴定概率，O 为蛋白质鉴定到肽段的期望，C 为样品中所有蛋白质的总量.	考虑蛋白质鉴定概率	[24]
PAI	$PAI_i=\frac{\#obsd_i}{\#obsbl_i}$	$\#obsd$ 为蛋白质鉴定到的肽段数目，$\#obsbl$ 为蛋白质理论可鉴定到的肽段数目.	考虑理论酶切肽段数目	[20]
emPAI	$emPAI_i=10^{PAI_i}-1$	------	指数变换改善线性	[25]
SIn	$SIn_i=SI_i/[\Sigma SI]/L_i$	SI 为蛋白质对应谱图的碎片离子强度总和.	消除重复实验的变化影响	[26]

1.2.1 基于谱图总数的校正方法.

相同的样品即使在同一实验平台下，不同批次实验产生的谱图总数的浮动也难以避免. 已有报道人类 T 白血病细胞蛋白质组技术重复之间的谱图总数差异达到 25%[27]. 为克服不同批次实验总谱图数变化带来的影响，最简单的校正方法就是依据各实验总谱图数进行归一化. 归一化谱图计数(normalized spectral counts，SpCn)模型即是用鉴定到蛋白质对应的总谱图数除以样品中所有蛋白质对应谱图数总和. Piersma 等[21]成功运用 SpCn 对 3 种不同的无标记定量蛋白质组学流程进行了比较，并对定量的重复性做了评估，3 种蛋白质组学流程均呈现较好的重复性. 在这类蛋白质组无标记定量分析中，前期实验步骤完全相同，比较定算结果才有意义.

1.2.2 基于蛋白质大小的校正方法.

最常见的谱图计数校正方法是根据蛋白质大小属性对原始谱图计数结果进行校正，其中包括蛋白质氨基酸序列长度[22, 28-29]、分子质量[30-31]和酶切肽段数目[20, 25]等. 蛋白质氨基酸序列越长，产生的酶切肽段可能越多，如此蛋白质长度给谱图计数定量带来一定偏性. 为此，人们通常将谱图数除以蛋白质长度作归一化处理.

NSAF(normalized spectral abundance factors)定义某鉴定到的蛋白质定量结果为该蛋白质的谱图数(SpC)除以该蛋白质的长度(L)，再除以样品中鉴定到所有蛋白质 SpC/L 比值的总和[22-23]. 研究表明 NSAF 定量动态范围可达 4 个数量级，能够检测低至 1.4 倍的丰度变化[22]. 另外 NSAF 也用于了哺乳动物介质复合体中不同蛋白质的丰度分析[32]，利用 NSAF 无标记定量方法，研究人员确定了一个常见的介质亚基的核心，几乎出现在所有细胞介质复合体中，然后，进一步用绝对定量(absolute quantification，AQUA)实验、定量蛋白质印迹以及转录组分析对 NSAF 结果进行了验证.

然而到目前为止，蛋白质分子质量与可鉴定到谱图数目的相关性仍不明确，有些研究表明蛋白质分子质量越大产生的谱图数会越多[33-34]. 但是 Lundgren 等[35]通过收集大量质谱数据分析却没有得到类似结论. 因此利用蛋白质分子质量大小对谱图计数归一化校正能否提高定量的准确性还需要进一步研究.

1.2.3 基于鉴定肽段属性的校正方法.

不同理化性质的肽段在质谱实验中会引入检测差异和偏性. 为克服肽段属性给谱图计数带来的弊端, Lu 等[24]开发了 APEX(absolute protein expression profiling)方法，引入了肽段的可检测性(肽段能被质谱检测到的概率). APEX 方法的关键在于一个校正因子 Oi，即能够检测到酶切肽段的期望，根据部分鉴定到肽段的长度和氨基酸构成等属性由机器学习算法计算得到. 这套方法基于一项肽段可检测性预测的研究[36]. 应用 APEX 的关键步骤就是如何选择合适的训练集来计算 Oi. 在离子阱质谱仪器上，APEX 的蛋白质定量动态范围达 3～4 个数量级. 另外研究人员还应用 APEX 估计了酵母和大肠杆菌中转录和翻译水平对基因调控的作用，定量分析结果与蛋白质免疫印记、流式细胞术、二维电泳以及转录组结果均一致. 目前已有免费的开源软件 APEX 定量蛋白质组学工具[37]. 在 2011 年 APEX 的研究被列为 *Nature Biotechnology* 杂志 5 年来引用最多的文章之一[38].

1.2.4 谱图计数方法的变型.

除了直接利用谱图计数方法，其他一些基于谱图计数思想并与之相类似的方法也已经广泛应用于无标记定量蛋白质组学，如利用肽段数目以及碎片离子信息等.

PAI(protein abundance index)模型用蛋白质鉴定到的肽段数目除以该蛋白质的理论酶切肽段数[20]作为该蛋白质的定量值，已有研究表明[25]PAI 定量结果与已知蛋白质丰度的对数值相关性较高(r=0.98). 同时，研究人员继续对 PAI 方法做出改进. 通过数学分析发现，蛋白质浓度与肽段数目的对数呈线性关系，再将参数进行指数变换后，线性关系进一步改善，于是诞生了 emPAI(exponentially modified protein abundance index)模型[25]，它与 PAI 模型的关系为 $emPAI=10^{PAI}-1$. 将每个蛋白质的 emPAI 值除以所有鉴定蛋白质 emPAI 总和作为该蛋白质的摩尔百分比含量，再用 BCA(bicinchoninic acid，二喹林甲酸)方法计算总共蛋白质含量，从而计算得出各蛋白质的绝对丰度. 其报道了小鼠全细胞裂解液中的 46 种蛋白质的绝对定量丰度与实际蛋白质浓度高度一致(r=0.93). 目前 emPAI 定量已经可以在 MASCOT[39]搜索结果中自动计算得出. 另外网页版工具 emPAI Calc[40]也可供免费使用，直

接输入 MASCOT 搜索结果文件即可．其他搜索引擎如 Sequest[41]和 X!TANDEM[42]等搜索结果文件格式转换后也可以使用 emPAI Calc．

在简单的谱图计数方法中，不管所含碎片离子强度的大小，所有谱图都计为“1”，这样导致谱图计数在定量低丰度蛋白质时存在明显不足．因此 Asara 等[43]开发了总离子计数(total ion count，TIC)方法来计算不同样品间蛋白质丰度的差异倍数，TIC 方法利用了蛋白质的所有二级谱图总碎片离子数目的平均值，提高了定量的线性动态范围．另外，为达到样品中蛋白质组分析的完整性，我们往往需要对该样品进行 4 到 8 次的重复质谱测量[44]．然而重复质谱数据含有固有的偏性和变化，为此 Griffin 等[26]开发了一种新无标记定量方法 SIn (normalized spectral index)，其整合三种质谱丰度特征，包括肽段数目、谱图数目以及二级碎片离子强度．经验证，与 NSAF 等其他谱图计数方法相比，SIn 大大消除了重复质谱测量之间的变化，并且通过免疫印记和密度测定方法进一步验证了 SIn 定量结果的正确性．与两种信号强度法方法[17, 45]的计算结果相比，SIn 可以更准确地计算出混合样品中蛋白质的含量．值得注意的是，作者使用的是低精度质谱仪器．有关高精度质谱仪器数据的性能比较有必要进一步研究．总之，SIn 强调了归一化定量的重要性，通过考虑蛋白质长度及所有碎片离子强度总和，消除了蛋白质长短和技术重复对信号强度变化的影响．

2 两类无标记定量方法的比较

2.1 质谱模式的选择

前面已经提到，在质谱仪器数据依赖获取模式下，进行一次母离子扫描后，会对选定的母离子进行二级碎裂，但是可用的色谱时间是固定的，因此质谱一级、二级扫描循环次数也是有限的．碎片离子扫描进行的越多，就会获取更多的二级谱图，肽段被鉴定到的可能性也就越大，进而也会增加鉴定蛋白质的数目[46-47]．然而，信号强度法为了准确计算蛋白质丰度往往需要利用更多的一级色谱峰，这样就会相对减少二级质谱的运行次数．可见，如果改变仪器运行模式优化信号强度法的定量性能就会相应减少鉴定蛋白质的数目．相反，谱图计数法依赖于鉴定肽段匹配的二级谱图数目，所以使谱图计数法发挥更好的性能就会增加鉴定蛋白质的数目[46]．然而需要注意的是，二级质谱运行的次数也不能过多增加，这取决于样品复杂程度、色谱峰宽度以及质谱信号获取速度等诸多因素．

信号强度法和谱图计数法对质谱运行模式的偏好不同，要求实验人员要均衡获取一级谱图与二级谱图，适当解决母离子信号强度法与蛋白质鉴定之间的矛盾．因此，一些实验室采用对每个样品执行两次独立质谱实验的方法，一次着重二级扫描，尽可能多地鉴定到肽段，另一次仅进行一级扫描以保证获取足够肽段母离子信号．此种方法缺陷在于需要结合准确的质量和保留时间来完成离子峰信号与鉴定肽段之间的匹配[48]．另外一种方法是采用数据独立获取(data-independent acquisition，DIA)模式[49]．DIA 模式不再采用一级扫描和二级扫描之间循环的模式，取而代之的是快速切换高、低碰撞能量．最近一项研究表明，DIA 获取方式可以提高信噪比 3～5 倍，鉴定到传统 DDA 模式下母离子扫描所检测不到的肽段[50]，在整个色谱峰范围内所有同位素峰同时提供母离子和碎片离子数据[48, 51]．这样，检测到的所有母离子信号强度都可以用于定量，既提高了可靠性又增加了动态范围．

2.2 数据处理复杂度

对于两类无标记定量方法的比较，数据计算方面不可忽略，其复杂程度直接决定着计算速度和效率．如前面所讲，信号强度法为了去除噪声信号与共洗脱成分信号往往要求极为复杂的平滑去噪、谱峰检测、保留时间对齐和归一化等处理，尤其在处理大规模实验数据时，这些步骤将会更加耗时．如何优化算法并提高计算速度将是信号强度法的主要发展方向之一．低精度质谱仪的背景噪声等通常更为明显，分辨能力也有限，应用信号强度法性能会大打折扣．随着质谱仪器精确性的增加，肽段质谱峰可以限定在更窄的质荷比范围内，使一些带有相似质量的干扰信号的影响降为最低，利于质谱峰检测[52]．因此就目前而言，信号强度法主要应用于高精度质谱仪器．相比较而言，谱图计数法主要基于一种经验性的观察，样品中某蛋白质的量越高，其对应的二级质谱谱图数越多，数据处理相对直观简洁，最简单的形式就是直接将蛋白质阳性鉴定对应的谱图数加和．只要前期实验处理部分流程一致，谱图计数结果就可以直接用来比较不同样品中蛋白质丰度的差异．

2.3　准确性和灵敏度比较

Xia 等[53]在差异蛋白质组学研究中发现谱图计数方法性能稍好于信号强度法，但是此研究注意到较小的谱图计数值往往造成很强的噪音，不适于进一步分析．尤其是当使用假发现率(false discovery rate，FDR)方法对质谱数据进行质量控制后，仅有少量蛋白质能够拥有较多数目的谱图．当蛋白质丰度很低时，肽段谱图计数可能为零，而肽段离子信号强度则可以获得更准确的结果．

Old 等[45]分析比较了信号强度法和谱图计数法的准确性和灵敏度．总体上，当蛋白质丰度相差大约 2 倍时，两类无标记定量方法均能区分，计算蛋白质丰度差异倍数结果基本一致．相比较而言，谱图计数法能够检测到更多差异蛋白质，也就是说灵敏度较高，而信号强度法则能更准确地计算蛋白质丰度差异倍数．Wienkoop 等[54]也验证了两种方法的线性和灵敏度，在差异蛋白质组学中的定量结果具有较好的一致性，但是复杂样品中定量结果的线性动态范围增加时，谱图计数可定量的蛋白质数目超过信号强度法．

Zybailov 等[55]利用离子阱质谱仪也发现谱图计数法在重复和动态范围性能上要优于信号强度法．最近，Dicker 等[56]又提出一种新方法 ProPCA，基于主成分分析的统计学方法整合二级谱图计数信息和一级质谱肽段离子峰强度信息．与基于谱图计数或仅基于肽段信号强度的方法相比，ProPCA 定量结果与已知蛋白质丰度取对数后相关性最强，并且在差异蛋白质组学研究中表现出了更好的灵敏度．ProPCA 的开发给无标记定量带来了新的视角，将信号强度与谱图计数整合，相互补充不足，为优化无标记定量提供新的方向．

3　无标记定量方法的优化

3.1　共享肽段的产生与处理

对于鸟枪蛋白质组学策略，在酶切肽段时，肽段和蛋白质之间的关联信息丢失，并且无标记定量信息获取初始阶段都是在计算肽段丰度．为获得蛋白质丰度，必须采取适当的方法将肽段的丰度分配至各蛋白质．因此，在鸟枪蛋白质组学中产生了“蛋白质组装问题”[57]．当一个肽段可以匹配至多个不同蛋白质序列时，我们称该肽段为共享肽段，如何决定共享肽段究竟来自于哪个蛋白质是一个难题．

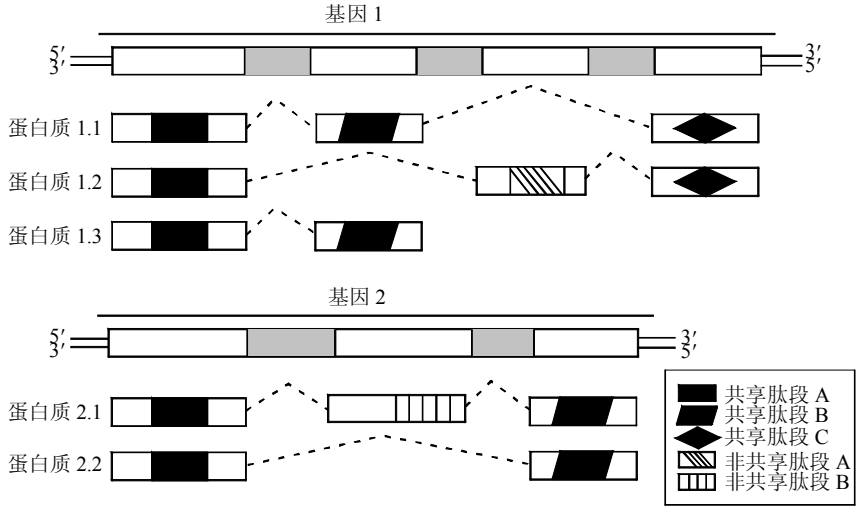

Fig. 2　Several sceneries for generation of shared peptides

图 2　共享肽段的产生

基因中灰色部分表示内含子，基因 1 产生 3 个可变剪切体，其中蛋白质 1.3 为蛋白质 1.1 的子蛋白．而基因 2 编码的蛋白质 2.2 与蛋白质 1.3 可能因同源关系序列完全一致．所有蛋白质中，仅蛋白质 1.2 和 2.1 具有非共享肽段，为可区分蛋白质，其他则难以区分．

3.1.1　共享肽段产生的原因．

肽段序列的鉴定依赖于二级谱图和蛋白质序列数据库的匹配比对．常用的蛋白质序列数据库很多，而且完整性和冗余度不尽相同．例如：NCBI

(National Center for Biotechnology Information) 的 Entrez Protein 数据库[58]在序列多态性和剪切体方面均较为完备，但是高度的冗余给蛋白质组装带来很大困难；同样来自 NCBI 的 RefSeq 序列数据库[59]则属于非冗余序列库，并具有详细的核酸序列和蛋白质序列关联信息；相比较而言，IPI(International Protein Index)数据库[60]在冗余度和完整性之间做到了很好的平衡．数据库的完整性保证更可能多地鉴定出肽段，但是蛋白质序列冗余度越高，共享肽段产生的可能性就越大．

另外，即使所用蛋白质数据库不存在冗余，一些生物学问题使共享肽段仍不可能避免．同源蛋白质、蛋白质亚型以及可变剪切体都会导致共享肽段．一个基因可能会转录翻译为成百上千个序列基本一致的蛋白质，极少量的短肽段可以用于区分这些蛋白质．甚至某些蛋白质序列只是另外一个蛋白质序列的一部分，它们的肽段全属于共享肽段，根本无法确定这些蛋白质是否存在．图 2 展示了一个产生共享肽段的例子，基因 1 和基因 2 可能是同源基因．其中仅有蛋白质 1.2 和 2.1 具有非共享肽段，当这两个非共享肽段得到鉴定时，则可以确认这两个蛋白质在样品中存在．其他蛋白质全由共享肽段构成，无法确定它们是否存在，即使确定其存在，怎样给它们分配肽段定量值也是难题．

3.1.2 共享肽段的处理方法．

处理共享肽段的一个最简单的方法就是忽略共享肽段，仅计算其他单一匹配肽段，这种方法的一个严重缺点就是低估了相关蛋白质的真实丰度[61]．还有人将共享肽段分配至共享该肽段的蛋白质群组[62]或者群组中的某一代表蛋白质[27, 63]．前者的蛋白质群组往往都是一些生物学相关的蛋白质家族成员，即使群组中的蛋白质之间丰度差异很大，这种方法也无法检测到；而后者根据的往往就是奥卡姆剃刀法则(Occam's razor)，用最少的蛋白质解释所有鉴定到的肽段．Feng 等[64]开发了程序 PANORAMICS，通过计算概率值来分配共享肽段．ProteinProphet[65]和 Mascot[66]等程序也是按照蛋白质的概率计算权重来分配蛋白质概率．目前较常用的分配共享肽段的策略是通过各蛋白质的单一匹配肽段丰度总和计算权重，将共享的肽段丰度值按比例分配至各蛋白质中[67-68]．Zhang 等[67]比较了多种计算权重的方法调整 NSAF 定量，最终发现根据单一匹配肽段丰度总和计算权重得到的结果与已知蛋白质丰度相关性最好；而 Fermin 等[68]利用这一策略开发了程序 Abacus，与未处理共享肽段方法相比，该方法计算两种同源蛋白质 BAF155 和 BAF170 的丰度差异更为显著．最近，有研究报道了一种更具逻辑性的新方法来处理共享肽段问题，首先利用一种贪婪集合覆盖算法(greedy set cover algorithm)生成最小蛋白质列表，对于含共享肽段的蛋白质群组，以各蛋白质单一匹配肽段值作距离进行聚类分析．生成的聚类图显示群组中蛋白质的关联，然后去除缺乏实验证据的蛋白质，这种分组聚类方法更容易在大规模数据中寻找目标蛋白质[69]．Qeli 等[70-71]则设计了 PeptideClassifier 方法，主张首先进行肽段水平的分类，根据蛋白质亚型以及在蛋白质和基因两种水平上是否单一匹配，将肽段划分成 6 类，生成最小蛋白质列表．PeptideClassifier 的优势在于详细考虑了基因、蛋白质序列以及蛋白质亚型 3 种水平的匹配信息，有助于靶向蛋白质组学和组学整合方面研究．

3.2 统计检验

前面提到的无标记定量方法除 APEX 和 emPAI 以外，大部分方法都是用于相对定量．如果需要获取某些蛋白质的绝对丰度往往需要在样品中加入标记的已知浓度的蛋白质[72]．除此之外，无标记定量方法最常用于检测不同样品之间蛋白质丰度的变化．然而只计算出蛋白质丰度的变化还不足以说明是样品间差异造成的丰度变化，还需排除系统的随机噪声等影响，因此丰度差异的显著性必须要进行统计检验，否则结论不能令人信服．通常，实验重复次数过少会阻碍丰度差异的统计检验，在有限的实验次数中，某些肽段对应的信号强度或是谱图计数结果都有可能产生零值，一方面会低估真实假阳性的数目，另一方面会产生丰度变化的极值，不能反映出真实的差异显著性．另外，实行参数统计检验的前提是数据呈现正态分布，而无标记定量数据由于离散特性常呈现出泊松分布，给统计检验带来一定困难．虽然统计检验方法也在不断推陈出新，可是统计检验方法的发展还是远落后于无标记定量方法的发展．

蛋白质组学相对是一个较新的研究领域，就其实验数据规模和性质来看，可以尝试借鉴基因芯片的统计检验方法．例如，LPE(Local Pooled Error)检验[73]和 PLGEM(Power Law Global Error Model)模型[74]就已经从芯片数据借鉴到谱图计数方法上[28]．Colinge 等[75]在包含 2～50 倍丰度差异蛋白质的样品中比较了 LPE 检验与 t 检验，结果发现在具有

2～3 次重复的实验中，LPE 检验优于 *t* 检验．然而，Zhang 等[76]却获得了不一致的结论，他们在含有 2 倍、5 倍和 10 倍丰度差异蛋白质的样品中比较了 5 种统计检验．其中包括 *t* 检验、LPE 检验、G 检验、AC 检验和 Fisher 精确检验，前两者必须要求 2 次以上重复．结果发现，在 3 次以上重复实验时，*t* 检验性能最佳，仅 2 次重复实验时，*t* 检验和 LPE 检验性能一般．利用 GLMM (generalized linear mixed effects model) 模型开发的统计软件 QSeq[29]不需要假定数据呈现正态分布，可以应用于自然泊松分布的数据，增加了使用灵活性．针对非正态分布的谱图计数数据出现了 ReSASC (resampling-based significance analysis for SpC) 工具[77]．最近 Pham 等[78]提出基于 β 双峰分布的检验方法，现已开发了 R 软件包可供下载使用．在与上述其他检验比较时，β 双峰分布检验在重复实验研究中性能最好．Van Breukelen 等[79]开发了 StatQuant 软件包，可以直接利用 MSQuant 质谱定量软件的输出结果，对信号强度无标记定量方法数据进行统计检验．

总之，在进行统计假设检验之前，需确定样品的重复实验次数，最好能结合多种统计检验方法的结果，得出的显著性差异结论更加可靠．

4 总结与展望

无标记定量方法的建立促进了定量蛋白质组学技术的发展．本文对常用的信号强度和谱图计数两种无标记定量方法进行较为全面的综述，并从对质谱仪器运行模式的偏好、数据处理复杂度以及灵敏度和准确性三个方面分析比较了两种方法的优缺点．需要我们注意的是，没有任何一种方法能够完美解决所有的问题，在进行选择信号强度法抑或是谱图计数法之前，需要考虑研究目的以及所用技术平台[80]．同时，本文详细讨论了无标记定量方法面临的两大难题，即共享肽段处理问题和统计检验问题，并列举了目前具有代表性的解决方法．

定量蛋白质组学的发展日新月异，无标记定量方法作为当前最常用的定量方法之一已得到广泛的应用．例如鉴定不同生物学过程中的表达谱[81]、诊断某些疾病[82]以及癌症[83]生物标志物、监测某些生物学过程蛋白质组的变化[84-85]以及蛋白质相互作用网络的研究等等[84]．像其他方法一样，无标记定量方法也存在缺点，对其优化改进也在不断进行，研究人员也尝试将无标记定量方法与其他方法结合使用．例如，Malmström 等[86]将无标记定量方法与多反应监测绝对定量结合使用，利用无标记定量数据覆盖度高的优势，将无标记定量结果转换为绝对定量结果，对生物学过程数学建模以及蛋白质相互作用网络等研究领域具有重要意义．随着质谱仪器性能的进一步提升和相关算法软件的开发，无标记定量方法定会不断完善，在今后的定量蛋白质组学领域占有一席之地．

参 考 文 献

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

[2] Ross P L, Huang Y N, Marchese J N, *et al*. Multiplexed protein quantitation in *Saccharomyces cerevisiae* using amine-reactive isobaric tagging reagents. Mol Cell Proteomics, 2004, **3**(12): 1154-1169

[3] Gygi S P, Rist B, Gerber S A, *et al*. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. Nat Biotechnol, 1999, **17**(10): 994-999

[4] Ong S E, Blagoev B, Kratchmarova I, *et al*. Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. Mol Cell Proteomics, 2002, **1**(5): 376-386

[5] Ryu S, Gallis B, Goo Y A, *et al*. Comparison of a label-free quantitative proteomic method based on peptide ion current area to the isotope coded affinity tag method. Cancer Inform, 2008, **6**: 243-255

[6] Domon B, Aebersold R. Options and considerations when selecting a quantitative proteomics strategy. Nature Biotechnology, 2010, **28**(7): 710-721

[7] Voyksner R D, Lee H. Investigating the use of an octupole ion guide for ion storage and high-pass mass filtering to improve the quantitative performance of electrospray ion trap mass spectrometry. Rapid Commun Mass Spectrom, 1999, **13**(14): 1427-1437

[8] 张 伟, 张纪阳, 刘 辉, 等. 蛋白质质谱分析的无标记定量算法研究进展. 生物化学与生物物理进展, 2011, **38**(6): 506-518
Zhang W, Zhang J Y, Liu H, *et al*. Prog Biochem Biophys, 2011, **38**(6): 506-518

[9] 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

[10] Wang G, Wu W W, Zeng W, *et al*. Label-free protein quantification using LC-coupled ion trap or FT mass spectrometry: Reproducibility, linearity, and application with complex proteomes. J Proteome Res, 2006, **5**(5): 1214-1223

[11] 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

[12] 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

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

[14] 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

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

[16] 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

[17] Silva J C, Gorenstein M V, Li G Z, *et al*. Absolute quantification of proteins by LCMSE: a virtue of parallel MS acquisition. Mol Cell Proteomics, 2006, **5**(1): 144-156

[18] Rauch A, Bellew M, Eng J, *et al*. Computational Proteomics Analysis System (CPAS): an extensible, open-source analytic system for evaluating and publishing proteomic data and high throughput biological experiments. J Proteome Res, 2006, **5** (1): 112-121

[19] Washburn M P, Wolters D, Yates J R, 3rd. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. Nat Biotechnol, 2001, **19**(3): 242-247

[20] Rappsilber J, Ryder U, Lamond A I, *et al*. Large-scale proteomic analysis of the human spliceosome. Genome Res, 2002, **12** (8): 1231-1245

[21] Piersma S R, Fiedler U, Span S, *et al*. Workflow comparison for label-free, quantitative secretome proteomics for cancer biomarker discovery: method evaluation, differential analysis, and verification in serum. J Proteome Res, 2010, **9**(4): 1913-1922

[22] 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

[23] Zybailov B L, Florens L, Washburn M P. Quantitative shotgun proteomics using a protease with broad specificity and normalized spectral abundance factors. Mol Biosyst, 2007, **3**(5): 354-360

[24] 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

[25] 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

[26] Griffin N M, Yu J, Long F, *et al*. Label-free, normalized quantification of complex mass spectrometry data for proteomic analysis. Nature Biotechnology, 2009, **28**(1): 83-89

[27] Wu L, Hwang S I, Rezaul K, *et al*. Global survey of human T leukemic cells by integrating proteomics and transcriptomics profiling. Mol Cell Proteomics, 2007, **6**(8): 1343-1353

[28] 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

[29] 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

[30] Powell D W, Weaver C M, Jennings J L, *et al*. Cluster analysis of mass spectrometry data reveals a novel component of SAGA. Mol Cell Biol, 2004, **24**(16): 7249-7259

[31] Peng J, Kim M J, Cheng D, *et al*. Semiquantitative proteomic analysis of rat forebrain postsynaptic density fractions by mass spectrometry. J Biol Chem, 2004, **279**(20): 21003-21011

[32] Paoletti A C, Parmely T J, Tomomori-Sato C, *et al*. Quantitative proteomic analysis of distinct mammalian Mediator complexes using normalized spectral abundance factors. Proc Natl Acad Sci USA, 2006, **103**(50): 18928-18933

[33] Cox B, Kislinger T, Emili A. Integrating gene and protein expression data: pattern analysis and profile mining. Methods, 2005, **35**(3): 303-314

[34] Gramolini A O, Kislinger T, Alikhani-Koopaei R, *et al*. Comparative proteomics profiling of a phospholamban mutant mouse model of dilated cardiomyopathy reveals progressive intracellular stress responses. Mol Cell Proteomics, 2008, **7** (3): 519-533

[35] Lundgren D H, Hwang S I, Wu L, *et al*. Role of spectral counting in quantitative proteomics. Expert Rev Proteomics, 2010, **7**(1): 39-53

[36] Mallick P, Schirle M, Chen S S, *et al*. Computational prediction of proteotypic peptides for quantitative proteomics. Nat Biotechnol, 2007, **25**(1): 125-131

[37] 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** (1): 529-539

[38] Baker M, DeFrancesco L. Five more years of Nature Biotechnology research. Nat Biotechnol, 2011, **29**(3): 221-227

[39] Pevzner P A, Mulyukov Z, Dancik V, *et al*. Efficiency of database search for identification of mutated and modified proteins *via* mass spectrometry. Genome Res, 2001, **11**(2): 290-299

[40] Shinoda K, Tomita M, Ishihama Y. emPAI Calc--for the estimation of protein abundance from large-scale identification data by liquid chromatography-tandem mass spectrometry. Bioinformatics, 2010, **26**(4): 576-577

[41] Yates J R, 3rd, Eng J K, McCormack A L, *et al*. Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. Anal Chem, 1995, **67**(8): 1426-1436

[42] Craig R, Beavis R C. TANDEM: matching proteins with tandem mass spectra. Bioinformatics, 2004, **20**(9): 1466-1467

[43] 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

[44] Li Y, Yu J, Wang Y, *et al*. Enhancing identifications of lipid-embedded proteins by mass spectrometry for improved mapping of

endothelial plasma membranes *in vivo*. Mol Cell Proteomics, 2009, **8**(6): 1219−1235

[45] 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

[46] Bantscheff M, Schirle M, Sweetman G, *et al*. Quantitative mass spectrometry in proteomics: a critical review. Analytical and Bioanalytical Chemistry, 2007, **389**(4): 1017−1031

[47] 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

[48] Silva J C, Denny R, Dorschel C A, *et al*. Quantitative proteomic analysis by accurate mass retention time pairs. Anal Chem, 2005, **77**(7): 2187−2200

[49] Venable J D, Dong M Q, Wohlschlegel J, *et al*. Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra. Nat Methods, 2004, **1**(1): 39−45

[50] Carvalho P C, Han X, Xu T, *et al*. XDIA: improving on the label-free data-independent analysis. Bioinformatics, 2010, **26**(6): 847−848

[51] Bateman R H, Carruthers R, Hoyes J B, *et al*. A novel precursor ion discovery method on a hybrid quadrupole orthogonal acceleration time-of-flight (Q-TOF) mass spectrometer for studying protein phosphorylation. J Am Soc Mass Spectrom, 2002, **13**(7): 792−803

[52] Hoehenwarter W, van Dongen J T, Wienkoop S, *et al*. A rapid approach for phenotype-screening and database independent detection of cSNP/protein polymorphism using mass accuracy precursor alignment. Proteomics, 2008, **8**(20): 4214−4225

[53] 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

[54] Wienkoop S, Larrainzar E, Niemann M, *et al*. Stable isotope-free quantitative shotgun proteomics combined with sample pattern recognition for rapid diagnostics. J Sep Sci, 2006, **29**(18): 2793−2801

[55] Zybailov B, Coleman M K, Florens L, *et al*. 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

[56] Dicker L, Lin X, Ivanov A R. Increased power for the analysis of label-free LC-MS/MS proteomics data by combining spectral counts and peptide peak attributes. Mol Cell Proteomics, 2010, **9**(12): 2704−2718

[57] 李 宁, 吴松峰, 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
Ling N, Wu S F, Zhu Y P, *et al*. Prog Biochem Biophys, 2009, **36**(6): 668−675

[58] Wheeler D L, Church D M, Edgar R, *et al*. Database resources of the National Center for Biotechnology Information: update. Nucleic Acids Res, 2004, **32**(Database issue): D35−40

[59] Pruitt K D, Tatusova T, Maglott D R. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res, 2005, **33** (Database issue): D501−504

[60] Kersey P J, Duarte J, Williams A, *et al*. The International Protein Index: an integrated database for proteomics experiments. Proteomics, 2004, **4**(7): 1985−1988

[61] 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

[62] 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

[63] Han M H, Hwang S I, Roy D B, *et al*. Proteomic analysis of active multiple sclerosis lesions reveals therapeutic targets. Nature, 2008, **451**(7182): 1076−1081

[64] Feng J, Naiman D Q, Cooper B. Probability model for assessing proteins assembled from peptide sequences inferred from tandem mass spectrometry data. Anal Chem, 2007, **79**(10): 3901−3911

[65] Nesvizhskii A I, Keller A, Kolker E, *et al*. A statistical model for identifying proteins by tandem mass spectrometry. Anal Chem, 2003, **75**(17): 4646−4658

[66] 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

[67] 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

[68] Fermin D, Basrur V, Yocum A K, *et al*. Abacus: a computational tool for extracting and pre-processing spectral count data for label-free quantitative proteomic analysis. Proteomics, 2011, **11**(7): 1340−1345

[69] Koskinen V R, Emery P A, Creasy D M, *et al*. Hierarchical clustering of shotgun proteomics data. Mol Cell Proteomics, 2011, **10**(6): M110 003822

[70] Grobei M A, Qeli E, Brunner E, *et al*. Deterministic protein inference for shotgun proteomics data provides new insights into Arabidopsis pollen development and function. Genome Research, 2009, **19**(10): 1786−1800

[71] Qeli E, Ahrens C H. PeptideClassifier for protein inference and targeted quantitative proteomics. Nat Biotechnol, 2010, **28**(7): 647−650

[72] Gerber S A, Rush J, Stemman O, *et al*. Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS. Proc Natl Acad Sci USA, 2003, **100**(12): 6940−6945

[73] Jain N, Thatte J, Braciale T, *et al*. Local-pooled-error test for identifying differentially expressed genes with a small number of replicated microarrays. Bioinformatics, 2003, **19**(15): 1945−1951

[74] 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**: 203−214

[75] Colinge J, Chiappe D, Lagache S, *et al*. Differential proteomics *via* probabilistic peptide identification scores. Anal Chem, 2005, **77**(2): 596-606.

[76] 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

[77] Little K M, Lee J K, Ley K. ReSASC: a resampling-based algorithm to determine differential protein expression from spectral count data. Proteomics, 2010, **10**(6): 1212-1222

[78] Pham T V, Piersma S R, Warmoes M, *et al*. On the beta-binomial model for analysis of spectral count data in label-free tandem mass spectrometry-based proteomics. Bioinformatics, 2010, **26**(3): 363-369

[79] van Breukelen B, van den Toorn H W, Drugan M M, *et al*. StatQuant: a post- quantification analysis toolbox for improving quantitative mass spectrometry. Bioinformatics, 2009, **25** (11): 1472-1473

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

[81] Duan X, Young R, Straubinger R M, *et al*. A straightforward and highly efficient precipitation/on-pellet digestion procedure coupled with a long gradient nano-LC separation and Orbitrap mass spectrometry for label-free expression profiling of the swine heart mitochondrial proteome. J Proteome Res, 2009, **8**(6): 2838-2850

[82] Sigdel T K, Kaushal A, Gritsenko M, *et al*. Shotgun proteomics identifies proteins specific for acute renal transplant rejection. Proteomics Clin Appl, 2010, **4**(1): 32-47

[83] Amon L M, Law W, Fitzgibbon M P, *et al*. Integrative proteomic analysis of serum and peritoneal fluids helps identify proteins that are up-regulated in serum of women with ovarian cancer. PLoS ONE, 2010, **5**(6): e11137

[84] Kim J, Choi Y S, Lim S, *et al*. Comparative analysis of the secretory proteome of human adipose stromal vascular fraction cells during adipogenesis. Proteomics, 2010, **10**(3): 394-405

[85] Sardiu M E, Cai Y, Jin J, *et al*. Probabilistic assembly of human protein interaction networks from label-free quantitative proteomics. Proc Natl Acad Sci USA, 2008, **105**(5): 1454-1459

[86] Malmström J, Beck M, Schmidt A, *et al*. Proteome-wide cellular protein concentrations of the human pathogen Leptospira interrogans. Nature, 2009, **460**(7256): 762-765

Label-free Methods in Quantitative Proteomics*

WU Peng, HE Fu-Chu, JIANG Ying**

(*State Key Laboratory of Proteomics, Beijing Proteome Research Center, Beijing Institute of Radiation Medicine, Beijing* 102206, China)

Abstract Mass spectrometry-based proteomics has developed rapidly, finding a method of quick, highly reproducible and accurate quantification is a great challenge in this research sphere. Quantitative proteomics as a new branch allows deeper insight into biological study. Recently, label-free quantification has been increasingly attractive for its simple sample preparation, low cost of reagents and clean results. Mass spectrometry-based label-free quantitative proteomics is generally divided into two categories, which are based on peptide chromatographic ion intensity and based on spectral counts. This review gives a general principle of these two label-free methods, presents a relatively complete summary of some commonly used label-free quantitative methods and their latest progress, and discusses the differences between the two methods and some challenges for label-free methods. We wish this review could provide a rational introduction of selecting label-free methods optimally suited to address your specific issue.

Key words quantitative proteomics, bioinformatics, label-free, signal intensity, spectral count, shared peptide
DOI: 10.3724/SP.J.1206.2012.00055

*This work was supported by a grant from National Basic Research Program of China (2010CB912700).

**Corresponding author.
Tel: 86-10-80727777-1213, E-mail: jiangying304@hotmail.com

Received: April 17, 2012 Accepted: June 1, 2012