中国科学: 生命科学 2015 年 第 45 卷 第 4 期: 347 ~ 358
SCIENTIA SINICA Vitae www.scichina.com life.scichina.com
《中国科学》杂志社
SCIENCE CHINA PRESS

评 述

基于质谱技术筛选差异表达蛋白的统计学策略研究进展

王锦霞[①], 常乘[②], 马洁[②], 吴松锋[②], 庄举娟[①*], 朱云平[②*]

① 大连海事大学数学系, 大连 116026;
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
* 联系人, E-mail: jjzhuang@dlmu.edu.cn; zhuyunping@gmail.com

收稿日期: 2014-07-17; 接受日期: 2014-11-27; 网络版发表日期: 2015-03-24
国家重点基础研究发展计划(批准号: 2013CB910800)、国家自然科学基金(批准号: 21105121, 21475150)、国家高技术研究发展计划(批准号: 2012AA020201)、蛋白质组信息学关键技术及分析系统研发国际合作项目(批准号: 2014DFB30010)和中央高校基本科研业务费(批准号: 3132015159)资助
doi: 10.1360/N052014-00197

摘要 随着质谱技术的快速发展, 蛋白质组学已成为继基因组学、转录组学之后的又一研究热点, 寻找可靠的差异表达蛋白对于生物标记物的发现至关重要. 因此, 如何准确、灵敏地筛选出差异蛋白已成为基于质谱的定量蛋白质组学的主要研究内容之一. 目前, 针对该问题的研究方法众多, 但这些方法策略的适用范围不尽相同. 总体来说, 基于质谱技术筛选差异蛋白的统计学策略可以分为 3 类: 基于经典统计学派的策略、基于贝叶斯学派的统计检验策略和其他策略, 这 3 类方法有各自的应用范围、特点及不足. 此外, 筛选过程还将产生部分假阳性结果, 可以采用其他方法对差异表达蛋白的质量进行控制, 以提高统计检验结果的可靠性.

关键词
质谱
蛋白质组学
差异表达蛋白
统计学原理
多重假设检验

蛋白质组学是后基因组时代兴起的一个重要的研究方向, 意在从整体水平上对组织或细胞内表达的全部蛋白进行定性和定量分析[1]. 蛋白定性分析起步较早, 随着质谱技术的不断发展, 已日渐成熟, 单一样本可以实现 8000 以上蛋白的鉴定规模[2]. 长久以来, 临床生物标记物的发现是蛋白质组学研究的热点, 对探索疾病机理和药物制备具有特别重要的意义, 而蛋白定量分析对这一研究的开展具有促进作用. 在定量研究方面, 围绕质谱数据进行差异蛋白筛选, 进一步实现生物标志物的发现与生物学特性的分析, 已成为定量蛋白质组学研究的一个重要方向.

基于质谱技术进行蛋白鉴定、定量及差异蛋白筛选的基本流程如图 1 所示, 可以分为实验和数据分析 2 部分. (ⅰ) 实验部分. 包括从生物样本的制备, 蛋白混合物的预处理, 到酶解肽段的质谱分析等一系列过程; (ⅱ) 数据分析部分. 包括质谱仪器获取原始数据后的所有数据处理的过程, 从蛋白鉴定、定量到差异蛋白筛选, 并且每一过程均涉及相应的质量控制和统计学分析. 基于质谱的定量研究可计算肽段

引用格式: 王锦霞, 常乘, 马洁, 等. 基于质谱技术筛选差异表达蛋白的统计学策略研究进展. 中国科学: 生命科学, 2015, 45: 347–358
Wang J X, Chang C, Ma J, et al. Statistical strategies for selection of differentially expressed proteins based on mass spectrometry technology. SCIENTIA SINICA Vitae, 2015, 45: 347–358, doi: 10.1360/N052014-00197

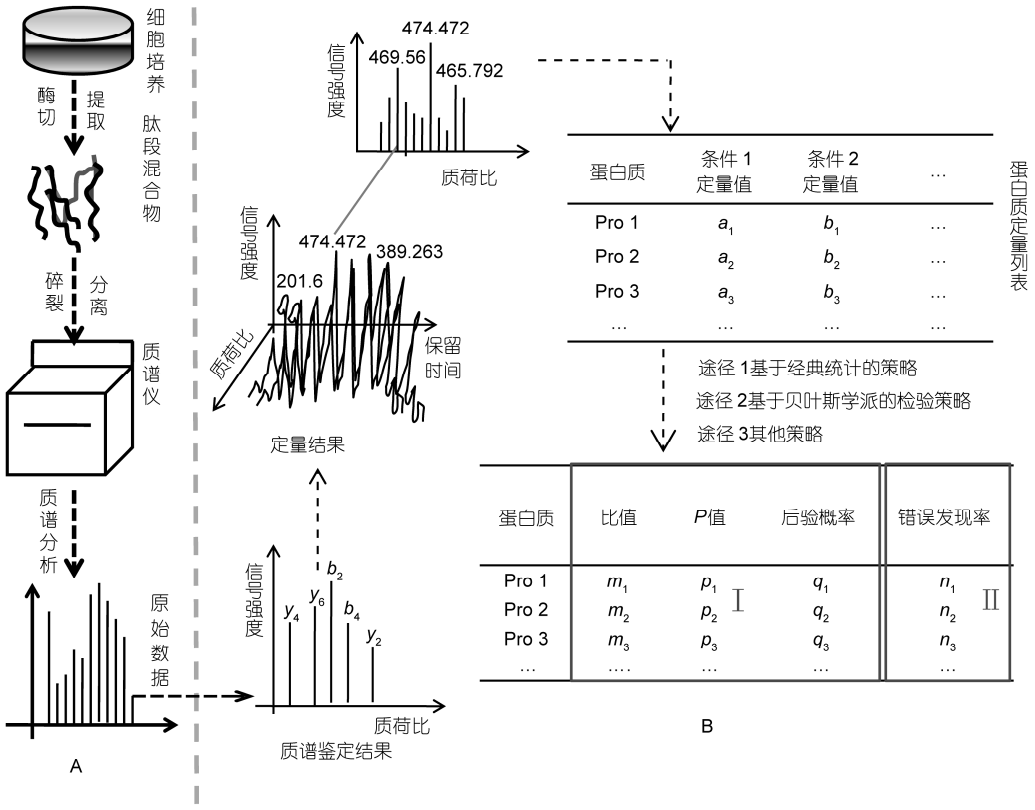

图 1　基于质谱技术的蛋白鉴定、定量及差异蛋白筛选的实验及数据分析流程示意图

A: 实验部分; B: 数据分析部分. Ⅰ: 将这 3 类值与已设定的阈值进行比较, 满足条件的蛋白便是差异表达蛋白; Ⅱ: 为保证候选生物标志物的可靠性, 需要进行质量控制, 降低数据分析结果的错误发现率

的丰度信息, 原则上, 可由肽段表达量推断出蛋白的表达量, 但此过程需要解决 2 个困难, 即肽段计算过程中信号峰缺失及共享肽段的问题. 对于前者, 可选择使用估计值填充缺失的数据, 以完善实验数据, 或者是在统计检验前, 利用肽段的天然同位素分布过滤噪声信号, 以选择最佳的肽段数据集进行实验[3]; 对于后者, 可选择使用合理分配的准则处理共享肽段. 肽段/蛋白表达水平值的准确定量对深入研究蛋白质组学意义匪浅, 它的一个主要目的是筛选差异表达蛋白. 差异表达蛋白指在不同实验条件下或不同的处理组中, 蛋白表达水平的检测值在排除系统随机噪声后达到一定的差异, 具有统计学意义, 同时也具有生物学意义, 差异蛋白筛选过程, 即是对定量结果做合理的统计推断. 因此, 利用统计学基本原理对差异表达蛋白进行显著性分析就显得十分重要.

总体来说, 定量蛋白质组学的数据分析存在 3 大主要问题: 数据的缺失值较多、实验的重复次数较少和结果的质量/可靠性参差不齐, 这对差异蛋白的筛选带来巨大挑战. 针对前 2 个问题, 研究人员已提出若干方法和工具, 这些方法各有利弊, 但很少能同时考虑 2 方面. 根据所属的统计学派别不同, 具体可以将它们分为 3 大类: 基于经典统计学派的策略、基于贝叶斯学派的统计检验策略和其他策略. 本文主要分析总结这些方法及工具的优缺点及应用范围. 针对第 3 个问题, 为了得到可靠的候选生物标记物, 可对统计检验筛选的结果进行质量控制, 这就需要在实验设计时利用基于内参的方法保证定量结果的可靠性, 或者是在多重假设检验中控制假阳性率. 而本文主要倾向于对后一种方法的介绍, 探讨了对筛选过程产生假阳性的控制方法, 最后还对目前研究中存在的问题以及未来的发展方向进行了讨论和展望.

1 差异蛋白筛选的统计分析策略

在理论发展的过程中, 统计推断衍生出 3 个主要的学派[4]: 经典学派、贝叶斯学派和信念学派. 在很长时间内, 经典学派占据着主流地位, 它的研究重点是样本空间上的概率分布, 并提出了大量影响深远的统计方法. 而在实际应用中, 人们逐渐应用贝叶斯学派的理念来处理问题, 它的研究重点是总体分布所处的状态空间. 信念学派的观点介于经典学派和贝叶斯学派之间, 自 20 世纪 60 年代以后, 针对它的研究越来越少, 所以本文将主要针对经典学派和贝叶斯学派的统计观点在筛选差异表达蛋白研究中的应用展开详细地叙述和讨论.

1.1 基于经典统计学派的差异蛋白筛选策略

根据是否要求实验数据服从特定分布, 可将经典学派的策略分为基于分布假设和基于传统非参数检验策略 2 类.

(1) 基于分布假设的统计策略. 基于分布假设策略的一般步骤可总结为: (ⅰ) 假设数据集服从某种特定的分布; (ⅱ) 建立统计模型、构造统计量; (ⅲ) 计算 P 值、确定阈值, 比较得出结论. 根据数据假设的分布不同, 又可以将方法细分为基于(对数)正态分布假设的统计策略和基于非正态分布假设的统计策略.

(ⅰ) 基于(对数)正态分布假设的统计策略. 研究某蛋白在 2 种状态下表达水平差异的显著性, 相当于检验 2 组数据的均值是否存在差异. 而 t 检验[5]是统计方法中发展较成熟的、用于分析 2 组样本间均值差异的方法, t 检验的前提假设为样本数据来自同一正态分布, 且要求每种样本至少进行 3 次重复实验. 当处理蛋白质组学数据时[6], 可选用肽段-电荷状态(peptide charge state, PCS)[7]代替自由度, 在肽段水平上计算蛋白的相对丰度(relative abundance)、加权合并的标准误差和相当于 t 检验的 P 值[8]. 但是, t 检验易受样本量的限制. 由于实验成本及时间等原因, 小样本的情况不可避免, 这就严重低估了总体方差, 导致结果中假阳性比例显著增加[9].

针对上述问题, 研究人员考虑对 t 检验进行改进, 减小由于统计量分母方差过小带来的错误. 2001 年, Tusher 等人[10]根据数据噪声大小与表达水平相关的特点, 对 t 检验进行修正, 提出芯片显著性分析(significance analysis of microarray, SAM)方法, 目的是筛选差异表达的基因. 由于蛋白质组学与基因组学有相似之处, 所以, Roxas 和 Li[6]将其借鉴到相对定量的质谱数据分析中, 且实验仅限于蛋白水平展开. 但是, 由于蛋白质组学的研究通常只进行单次或 2 次重复实验, SAM 方法和 t 检验均需要至少 3 次重复, 所以, 对蛋白质组学的数据来说, 2 种方法的准确性较低.

同样, 为了降低由于 t 统计量分母过小而引发的高假阳性率, Jain 等人[11]提出 LPE 检验(local pooled error test). 2004 年, Allet 等人[12]将其引入不成对蛋白质组学数据的研究中, 特别是用于分析基于谱图计数(spectral count, SC)的定量结果. 在检验时先合并具有相似表达水平蛋白的方差, 然后将 2 个状态下表达水平中位数间的差异与合并方差的比作为 z 统计量, 与 LPE 检验在基因组中的应用相比, 此实验要求至少进行 2 次重复. 2007 年, Cho 等人[13]对 LPE 方法重新定义, 利用成对标记质谱数据进行实验. 修改后的 LPE 方法采用秩不变量重新抽样技术(rank-invariant resampling technology)处理不服从正态分布的数据, 并允许缺失值的存在, 此方法为解决定量蛋白质组学研究中缺失值较多的问题提供了一个可选择的方案, 但对于只有 1 次重复的蛋白质组学数据来说, 它的实验结果会包含较多的假阳性.

当研究某蛋白在至少 3 种状态下表达水平的差异时, 可利用方差分析(analysis of variance, ANOVA)方法, 又称变异数分析或 F 检验. 但是此方法只能筛选出哪个蛋白存在差异, 而不能区分差异属于哪些状态之间, 所以一般要增加两两比较(post-hoc comparisons)检验作为补充. 这就涉及多重假设检验问题, 在本文的后续部分会进行详细介绍.

Clough 等人[14]认为, 蛋白定量和显著性分析的准确性直接影响系统生物学的研究和生物标记物的发现. 针对此问题, 它们提出一个线性混合效应模型, 假设随机误差项服从正态分布. 对于有限的实验重复次数和较多的缺失值问题, 此方法要求实验在足够多的状态下进行. 为推广应用, 这一方法可以通过基于 R 的开源软件包 MSstats 实现.

(ⅱ) 基于非(对数)正态分布假设的统计策略. 由于肽段的谱图计数一般表现为离散特征, 经对数变换后也并非近似正态分布, 而泊松分布恰好可以描述这一类数据, 但它要求数据满足“均值=方差”[15],

这一前提对定量蛋白质组数据的经验拟合并不灵活, 特别是基于谱图计数的定量数据会产生“过离散”现象(即方差大于均值)[16].

为解决这一问题, Li 等人[17]基于最大似然, 在回归框架中对谱图数建立泊松类似然模型. 为比较 2 个或多个状态下蛋白的表达差异性, 在允许计数数值存在极端离散的情况下, 模型假设数据满足“方差和均值呈线性相关”的特点, 这有效降低了泊松模型的假设要求. 此模型的优点在于无需数据的任何先验信息就可以估计蛋白的平均谱图数, 但不能解决实际数据中的“低方差”和“缺失值”问题.

在处理离散数据方面, 可以和泊松类似然模型相媲美的是负二项模型[16], 用于 2 个独立样本间的检验. 它假设数据服从负二项分布, 且方差是均值的二次函数, 此模型可处理“小样本”和“异方差”问题, 对“过离散”的数据来说是一个很好的预测器, 预计将在定量蛋白质组研究, 特别是在基于谱图计数定量结果的研究中发挥重要作用. 贝塔二项(beta-binomial, BB)模型[18]同样可用于 2 独立样本间的检验, 此方法的最大特点在于同时考虑样本的组内方差和组间方差. 而为进行成对样本间的差异分析, Pham 和 Jimenez[19]在 BB 模型的前提下, 提出 inverted Beta-Binomial(iBB)模型, 这种方法不仅可用于蛋白质组学数据, 也可用于基因组学的数据分析中.

(2) 基于传统的非参数统计检验策略. 基于分布假设统计策略模型均有一个明显的不足: 检验时需要假设数据服从某种分布形式. 虽然基于质谱的蛋白质组学产出的是大规模数据, 在理论上会趋近一些常用分布, 但在实际应用中, 总会有某种分布难以描述的情况存在. 而非参数检验方法是直接对统计量的分布进行估计, 不要求数据满足特殊分布, 在这一点上优于上述策略, 也能更好地分析蛋白质组学数据.

Fisher 精确检验[20]、G 检验[21]和 Audic-Claverie (AC)检验[22]是 3 个较早提出的非参数检验方法, 适合分析实验重复次数有限的数据, 这也恰好满足蛋白质组学研究中“少重复”的特点. 其中, Fisher 精确检验可用于肽段计数资料的显著性分析[23]. AC 检验的关键是计算条件概率: $P(x_i \mid x_j)$ (其中, x_i 和 x_j 分别指在实验条件 i 和 j 下, 观察到的谱图数, i, j=1, 2). G 检验是卡方独立性检验的修改版, 可分析存在缺失值的数据.

通常, 多个状态下的差异分析相对于 2 个状态下的差异分析更复杂. 在不要求进行重复实验的情况下, Zhang 等人[24]扩展了 G 检验, 使构造的抽样统计量服从卡方分布并计算 P 值, 通过 Benjamini-Hochberg(BH) 理论[25]进行校正, 根据不等式 $P \leqslant (j/t)q$(其中, t 代表被检验的蛋白总数, j 指将所有蛋白的 P 值按升序排列后的第 j 个位置, q 代表指定的错误发现率(false discovery rate, FDR)[25])就可对 6 种不同状态下差异表达蛋白进行筛选. 出于 G 检验的这一特点, Cooper 等人[26]用它分析来自 9 次重复实验的 MudPIT(multidimensional protein identification technology)[27]样本数据, 并对离子抽样误差与谱图计数间的关系进行相关研究. G 检验能较好地处理“多缺失、少重复”的蛋白质组学数据, 可考虑在以后的研究中进一步推广应用.

另外一种常用的非参数假设检验为 Wilcoxon 秩和检验(Wilcoxon rank sum test)[28], 又称 Mann-Whitney 检验. 此检验只需利用秩的信息构造统计量, 异常数据对它的影响较小, 但采用该方法的检验功效较低, 会导致某些蛋白的统计量值相同无法实现进一步区分, 建议结合其他的方法共同使用.

1.2 基于贝叶斯学派的统计检验策略

不同于经典统计学的方法, 贝叶斯学派针对统计检验具有独特的流程, 可简单总结为: (ⅰ) 获得数据的先验信息; (ⅱ) 由贝叶斯法则计算后验概率; (ⅲ) 利用贝叶斯因子(实质上是 2 个模型的似然比)判断 2 个假设的可靠性.

Efron[29]利用贝叶斯学派的观点提出的经验贝叶斯模型, 能够克服高斯混合模型的局限性, 可用于处理有标定量的蛋白质组学数据, 但它对非高斯截尾或存在缺失值的数据并不适用. 并且, 此模型只是在理论上较好地拟合原假设, 对于真实数据, 这个方法的拟合结果并不是很理想. 因此, 实验得到的后验概率便不准确, 从而导致假阳性结果的产生. 所以, Margolin 等人[30]在 SILAC(stable isotope labeling by amino acids in cell culture)[31]定量蛋白质组实验数据的基础上对此模型进行修改, 进一步拓展其在蛋白质组学中的应用——不仅能鉴定小分子的目标蛋白, 还可判断是否由微小核糖核酸导致蛋白的差异调控, 并且, 提出的多元统计量可用于计算差异表达蛋白

的数目、统计功效和错误发现率.

为提高差异表达蛋白的检测率, Koopmans 等人[32]在经验贝叶斯理论的基础上提出 EBRCT(empirical Bayesian random censoring threshold)模型, 此方法可以在肽段和蛋白水平上, 对带有缺失值且重复次数为 6 次的样本进行显著性分析. 并且, 实验证实 EBRCT 模型优于基于置换的 2 个模型: IKNN(K-nearest neighbors)[33]和 SVTI(singular value thresholding)[34].

Choi 等人[35]在泊松模型的基础上, 基于分层贝叶斯估计提出统计学分析框架——QSpec, 将谱图数看做随机数, 把贝叶斯因子当做判断统计显著性的指标, 可以直观地判断零假设/备择假设 2 种模型中哪种能更好地解释实验数据, 而不是简单地“接受”或“拒绝”, 此方法可用于任意重复次数的实验中, 并且较适合处理基于谱图计数的定量数据. 与此类似, Booth 等人[36]基于谱图数构建贝叶斯混合模型, 定义与似然比检验统计量相似的 BF-统计量, 在 H_0 条件下, 用统计量的累积分布密度函数代替假设检验的 P 值, 通过伪贝叶斯因子筛选差异表达蛋白.

Serang 等人[37]基于 Dirichlet 过程提出 npCI(non-parametric cutout index)方法, 通过比较不同蛋白数据集中 remaining PSM(peptide-spectrum match)与 empirical absent PSM 的打分分布, 赋予它们不同的似然比, 它是利用非参数贝叶斯拟合评价其他统计方法的一种非参数统计方法. 同年, Serang 等人[38]将此方法推广到定量蛋白中, 用以评估鉴定差异表达蛋白的统计策略.

1.3 其他策略

事实上, 已提出的针对差异蛋白筛选的统计方法, 并不能严格地按照统计学派划分, 例如, 经验贝叶斯方法, 就是经典学派方法与贝叶斯学派观点的有机结合, 直到今天都影响深远. 因此, 除上述基于两大统计学派发展差异蛋白筛选策略外, 还存在一些其他策略. 它们或者是从其他组学方法中借鉴, 或者是综合定量和定性 2 种分析技术, 但在方法的应用过程中, 同样都离不开统计学理论的框架.

蛋白质组相对于转录组和基因组起步较晚, 从实验数据的大小、性质和实验目的来看, 可以尝试借鉴其他组学的检验方法. 但是, 由于蛋白质组学数据有自己的特点, 因此在引用方法的过程中需注意根据实际进行优化. 最初用于 cDNA 微阵列差异研究的方法是倍数分析(fold change, FC), 通过比较某基因在 2 个状态下表达水平的比值(即 ratio 值)和所选阈值, 就可得到差异表达基因. 将此方法引入蛋白质组时, ratio 代表 2 个条件下某蛋白丰度的比. FC 方法简单方便, 能够节省大量的实验成本, 并且对没有重复实验的数据仍然适用, 这也是其在生物标记物临床研究中广泛应用的原因. 虽然实验结果具有生物学意义, 但由于此方法不能指出差异蛋白内在的置信水平, 所以得到的实验结果不具有较强的统计学意义, 一般不推荐单独使用[6].

为了在多次重复的实验中鉴定差异表达的基因, 可以综合使用 PLGEM(a power law global error model)模型[39]和信噪比(signal-to-noise ratio, STN)检验统计量. 考虑到随机噪声对蛋白丰度的影响, Pavelka 等人[40]用标准化的谱丰度因子(normalized spectral abundance factor, NSAF)对蛋白表达水平的观测值进行归一化处理, 以减小数据偏差, 巧妙地将这一基因组学中的方法引申到基于谱图计数定量结果的分析中, 且验证了 NSAF 与转录组丰度值有相似的统计属性, 但实验最终并没有给出相应的显著性度量指标(如 P 值). 针对低丰度蛋白的差异分析, 与 FC 和 Standard-STN 方法相比, PLGEM-STN 方法更保守. 为进一步拓展 PLGEM-STN 在定量蛋白质组学中的应用, Li 和 Roxas[41]将其与 MPSP(minimum number of permuted significant pairings)规则结合, 共同分析基于信号强度的定量数据, 该实验可以处理只有单个肽段匹配或倍数变化小的蛋白, 与单独使用 PLGEM-STN 方法相比, 本实验方法的分辨率有所提高, 且在没有显著影响灵敏度的前提下, 有效减少结果的假阳性和阳性数目, 保证了差异蛋白的可靠性.

以上所有方法及工具都是在定量结果的基础上提出的, 但是定量的过程通常会受到缺失值的影响, 肽段/蛋白若以低丰度的形式存在, 则很难被鉴定到, 这就对定量蛋白质组学的研究造成巨大困难. 所以, 若仅应用定量差异分析技术, 产生的结果不一定可靠[42]. 目前常用的定性分析技术包括 bottom-up 和 top-down 路线, 它们只能提供蛋白有/无的信息, 当此信息的含量低于体系检出下限时, 则判断为无, 对于样品中都能检出的蛋白不能提供丰度差异的信息[43]. 如果综合使用定量和定性 2 种分析技术, 定会增加实验结果的可靠性.

因此, 2009 年, Karpievitch 等人[44]提出基于 bottom-up 路线的统计学框架, 此框架能在肽段信号峰大量缺失的情况下, 对蛋白的表达水平进行无偏估计, 在进行统计推断时, 构造似然比检验统计量, 其分布函数的截尾概率相当于 *P* 值. 2010 年, Webb-Robertson 等人[42]提出 IMD-ANOVA (independence of missing data with an analysis of variance)方法, 利用定量、定性 2 个统计置信指标(峰强度、检测结果)表征肽段信息, 将 ANOVA 方法和 G 检验结合, 在肽段水平上分析蛋白质组学数据, 并定义检验统计量为组间方差和组内方差的比值, 通过 *F* 检验计算 *P* 值. 2012 年, Wang 等人[45]根据峰的可检测性将强度转化成二进制数, 这和谱图计数法类似, 并基于二项式似然开发出一个混合的统计策略, 包括基于强度的定量分析和存在/缺失分析, 每次分析均会产生 1 个差异蛋白列表, 实验最终得到的是所有列表混合后的单列表. 上述 2 个实验策略的基本思想是相似的, 结果均可保留既存在定量、又存在定性差异的蛋白, 这对进一步挖掘可靠的生物标记物提供了很大帮助.

表 1 总结了上述所列举的 3 大类统计方法的优缺点、适用范围及有代表性的方法. 由于这些方法适用的数据类型不尽相同, 有的是基于谱图计数定量结果, 有的是基于信号强度定量结果, 还有的是基于有标定量的定量结果, 并且, 针对蛋白质组学数据分析中“缺失值多”、“重复次数少”的 2 大问题, 它们的解决力度也不相同. 因此, 暂时没有合适的实验数据或模拟数据用来评估全部的方法.

2 差异蛋白筛选过程中的假阳性控制问题

在蛋白质组学的差异分析中, 检验结果呈假阳性(false positive, FP)的表现为, 某些蛋白已知正常表达却误判为差异表达. 满足这样条件的蛋白在假设检验中称作假阳性结果, 其所占检验结果中阴性蛋白总数(*N*)的比例就是假阳性率(false positive rate, *FPR=FP/*(*FP+TN*)). 对于假设检验问题, 不论选用哪种检验方法, 都存在犯 2 类错误的概率, 特别是由于蛋白质组学数据的复杂性和统计策略的不完备性, 筛选结果的质量控制问题十分突出. 为了使筛选出的差异蛋白更令人信服, 就必须控制检验结果的假阳性率. 表 2 描述了假设检验前后蛋白差异性变化的关系.

对于只有 1 个蛋白的情形, 处理方法较简单, 只需进行单个假设检验. 过程大致分为 5 步: (ⅰ) 确定原假设和备择假设; (ⅱ) 构造检验统计量; (ⅲ) 确定检验水平 α; (ⅳ) 计算检验统计量的 *P* 值; (ⅴ) 将所得的 *P* 值与 α 比较, 看是否满足 $P\leqslant\alpha$. 一般设定 α=0.05 为差异表达, α=0.01 为显著差异表达.

但是, 通过高通量的质谱技术, 研究者能够在 1 次实验中得到大量的定量结果. 现有基于统计学原理筛选差异蛋白的方法和工具, 均会产生或高或低的假阳性率. 若仍利用单个假设检验中的 *P* 值作为判断标准, 则会导致结果假阳性率的累积. 例如, 在显著性水平为 0.01 的情况下, 若同时对 10000 个蛋白进行假设检验, 则最终会有 100 个蛋白呈假阳性. 对于这种情况, 常用的解决办法是: 在多重假设检验中对 *P* 值进行校正, 进而实现假阳性率和假阴性率(false negative rate, *FNR=FN/*(*FN+TP*))的平衡[47].

多重假设检验(multiple hypothesis testing)是指同时对多个假设进行检验, 首先将多个单重的假设检验作为一个整体, 然后对这个整体中的所有假设同时进行检验. 可以简单地理解为, 1 次数据分析中包含的蛋白总数(*M=TP+FN+FP+TN*)就是假设检验的重数.

2.1 多重假设检验的校验准则

进行多重检验时, 最重要的是控制整体错误率, 这就需要某种准则将错误控制在一定范围之内. 广泛使用过的错误度量指标是错误率判断族(family wise error rate, FWER)[48]、错误发现率(FDR)[25]和正错误发现率(positive false discovery rate, pFDR)[49].

FWER 准则的定义为 *FWER=Pr*(*TP*⩾1), 表示检验结果中至少出现 1 次假阳性的概率. 对此准则通常使用的校验方法是邦弗朗尼(Bonferroni)法[50]和 Westfall and Young 逐步向下校验法[5]. 两者均认为不同的假设检验间是独立的, 但后者较严格, 因此挑选出的差异蛋白数目较少. 总之, *FWER* 准则允许实验结果包含一定的假阳性, 在同方差模型下, 通过基于重复抽样的 Permutation 和 Bootstrap 进行控制. 但是它太保守, 不适合高通量、大范围的同时假设检验问题[17], 这就需要寻找一个替代指标.

1995 年, Benjamini 和 Hochberg[25]提出新的度量准则 *FDR*, 又称 BH *FDR*, 定义为: *FDR=FP/*(*FP+TP*),

表 1 统计方法的分类、优缺点及代表性方法

方法的分类		优点	缺点	有代表性的方法	适用范围		参考文献
基于经典统计学派的方法	基于正态分布	简单方便, 按照多重假设检验的步骤依次计算统计量和 *P* 值, 并进行检验结果的校正	当数据不服从(对数)正态分布时, 方法不可用	*t* 检验	至少 3 次重复	信号强度、连续化的谱图计数、不允许存在缺失值	[5]
				ANOVA			[6,10]
				SAM 方法			
				LPE 方法	至少 2 次重复		[11~13]
	基于非正态分布	可处理“过离散”的数据	当数据不服从任何分布形式时, 方法不可用	负二项模型	无缺失值	谱图计数、不要求重复实验的次数	[16]
				泊松类似然模型			[17]
				负二项广义线性模型	允许存在缺失值		[46]
	基于非参数统计	不要求数据满足特殊分布形式, 直接对统计量的分布进行估计	检验的功效较低	Fisher 精确检验	至少有 1 组无缺失值	谱图计数、1 次或 2 次重复试验	[20,23]
				AC 检验			[22]
				G 检验	无缺失值		[21,24,26]
				Wilcoxon 秩和检验	至少 3 次重复, 允许存在缺失值, 但不可 1 组内全缺失	谱图计数、信号强度	[28]
基于贝叶斯学派的方法		统计指标变为贝叶斯因子, 可直观地判断零假设/备择假设 2 种模型中哪种能更好地解释实验数据, 而不像经典统计学派中简单地“接受”或“拒绝”	数据的先验信息有时不可知, 后验概率的计算较麻烦	经验贝叶斯模型	无缺失值	SILAC	[29,30]
				EBRCT 模型	允许存在缺失值, 对重复次数无严格要求	信号强度	[32]
				贝叶斯混合模型	无缺失值	对重复次数无严格要求 谱图计数	[36]
				npCI 模型	允许存在缺失值	对重复次数无严格要求 PSM	[37,38]
其他方法		可将其他组学的方法引申到蛋白质组学中, 使得方法在组学数据的分析中具有通用性; 也可将定性差异分析与定量差异分析相结合, 最大化差异蛋白的数目	蛋白质组学的数据有独特的特点, 在方法的引申时只能借鉴, 不可完全照搬; 定性差异分析方法的发展还有待完善	倍数分析	允许存在缺失值, 对重复次数无严格要求	谱图计数、信号强度	
				PLGEM 模型	1 次或 2 次重复	NSAF	[39]
				IMD-ANOVA 方法	允许存在缺失值, 对重复次数无确定要求	信号强度、连续化的谱图计数	[42]
				混合统计方法			[45]

它实际是一个估计的假阳性率. 在差异蛋白筛选的问题中, *FDR* 与 *FPR* 的定义虽不同, 却有相似的生物学意义, *FDR* 表示阳性检验结果中判断错误的比例, 是一个基于频率的方法. *FWER* 主要是控制第一类错误, 相对而言, *FDR* 可为总阳性率和假阳性率之间提供一个较好的平衡. 换言之, *FDR* 在有效控制假阳性结果的同时, 能最大化差异表达蛋白的数目, 所以它可以作为所需要的指标.

根据上述 *FDR* 的定义可以发现, 当实验结果的阳性总数为 0(即经统计分析不存在差异表达蛋白)时, *FDR* 并不能合理地说明问题, 因此研究者便对其进行修改得到 *pFDR*, 其定义为 $pFDR=E[(FP/(FP+TP))|(TP+FP>0)]$, 比较 *pFDR* 和 *FDR* 两者的定义可以得到下面的关系式: $FDR=pFDR\cdot Pr(TP+FP>0)$, 所

表 2　假设检验前后蛋白差异性变化的关系

		假设检验的结果		合计
		差异表达	正常表达	
已知条件	差异表达	*TP*(true positive)	*FN*(false negative)	*TP+FN=P*
	正常表达	*FP*(false positive)	*TN*(true negative)	*FP+TN=N*
合计		*TP+FP*	*FN+TN*	*TP+FN+FP+TN=M*

以在假阳性结果的控制方面, 2 个准则可以互相借鉴.

2.2　假阳性结果的控制

通过对 3 个准则的简单介绍, 本小节以 FDR 准则为代表讲述假阳性结果的控制问题. 而控制结果的假阳性数目, 实质上就是控制 *FDR*, 所以在多重检验中正确估计错误发现率, 并进行有效控制, 对高通量数据的显著性分析来说十分重要.

FDR 的估计与控制. *FDR* 的估计是指在某一显著性水平下假发现率的估计. 针对 *FDR* 的估计问题, 除利用定义直接计算外, 研究人员还提出一些其他的解决方法. 例如, 2007 年, Cho 等人[13]用原始数据和重新抽取的无效数据集计算检验统计量, 并与给定的 Δ 值比较, 以此判断肽段的显著性, 将 *FDR* 定义为 $FDR(\Delta)=\alpha\cdot V_1/V_2$ (其中, α 为校正因子, V_1 为无效数据中差异表达肽段数, V_2 表示原始数据中差异表达肽段数); 2013 年, Fu 和 Qian[51]针对翻译后修饰结果的质量控制问题, 利用特定搜库策略获得经验数据, 计算局部错误发现率(local FDR)和子集错误发现率(subgroup FDR)的理论关系, 直接从整体上估计最终数据集的 *FDR*, 避免从有限数据中估计带来的麻烦, 当只有很少的修饰被鉴定到时, 此方法优于 target-decoy 搜索策略[52]; 2014 年, Tan 和 Xu[53]基于 BH 理论开发出一个新算法, 功能与 SAM 类似, 只需指定 *C*-value(表示控制值), 算法就可以在 BH 程序、Bonferroni 程序和单个假设检验程序间做出选择, 利用多项式回归方法计算 *FDR*, 虽然在决定参数 *C* 时耗时较长, 但 *FDR* 的估计值较 SAM 更接近真实值.

以上均是从经典统计的角度估计 *FDR*, 而贝叶斯检验的后验概率反映了真正的期望错误率, 检验标准是接受具有最大后验概率的假设, 且更易推广到多重假设检验的场合. 所以还可考虑从贝叶斯角度对 *FDR* 给出解释, 在此框架下也提出了相应的估计方法. 例如, Farcomeni[54]利用两成分模型计算 *P* 值的分布函数, 从而将 *FDR* 定义为 $FDR=P_0F_0(P)/F(P)$ (其中, $F(P)=P_0F_0(P)+P_1F_1(P)$, 表示实验结果中表现差异的蛋白所对应 *P* 值的分布函数, $P_0=FP/M$).

FDR 的控制是指决定 1 个显著性水平 α 的阈值, 使 *FDR* 限制在某一固定水平. 已提出的控制策略中, 大部分是设置统计量的阈值. 而 Efron[55]认为, *FDR* 的控制问题就是一个贝叶斯问题, 且 BH 方法、Storey 方法[56]只是其中的特例. 因此, 在经验贝叶斯模型中, Efron 将贝叶斯后验概率的阈值设为 0.8 来控制 *FDR*[29]. 虽然经验贝叶斯方法在理论上可以对原假设进行较好地拟合, 但对实际数据并不能得到理想的结果. 为此, Bei 和 Hong[57]提出一种新的 *FDR* 控制方法, 称作 *miFDR*. 在事先给定所需的显著特征数目的前提下, *miFDR* 可对 *FDR* 进行最优化处理. 实验将此方法与 BH 方法、Storey 方法和 SAM 进行比较, 结果显示, 在相同的 *FDR* 阈值下, *miFDR* 可鉴定更多的显著特征. 当不考虑灵敏度时, Li 和 Roxas[41]利用 3 个特定的参数(也就是倍数分析的阈值、某种检验统计量的阈值和 MPSP 规则)共同控制 *FDR*, 此方法在不需进行重复实验的前提下, 既能评估统计显著性, 也能提供较高的分辨率.

统计显著性指标除上述提到的 ratio 值、*P* 值、后验概率和 *FDR* 外, 还有 *q* 值[49], 它是所有 *FDR* 中最小的 1 个. 对其中部分指标, 研究人员已进行了相关的比较分析[58~60], 本文不再重复评估. 由于本文的重点在于综述筛选差异表达蛋白的统计方法, 所以对于检验结果的质量控制问题着重阐述了多重假设检验方法. 但是, 在质谱领域还存在 1 种基于内参的质控策略, 就是将内参样本(作为阳性)加入已准备的样本(作为阴性)中, 以便评估筛选差异表达蛋白的方法, 目前很多研究人员在使用这种基于内标的实验设计[6,12,17,39,61].

纵观上述筛选差异表达蛋白的统计策略, 它们有一个共同的特点: 考虑的问题不够全面. 现有的方法要么只适应“多缺失”的数据, 要么只适应“少重复”的数据, 同时满足 2 个条件的方法较少, 并且, 在进行数据分析时, 都需要严格的多重假设检验来控制假阳性率. 因此, 针对蛋白质组学数据分析的 3 大问题, 目前并没有十分完美的解决方案. 随着各种计划的开展, 单次实验的数据量显著增加, 对于统计方法而言, 计算变得更加复杂, 针对这种情况, 需要改进现有方法, 使它们继续保持好的检验功效. 例如, *t* 检

验适用于小样本数据, 当样本量大于 50 时, 可选择 U 检验, 且正态性要求可以放宽. 所以, 在理论上, 应该能找到代替现有的其他方法的更优策略.

3 差异蛋白筛选的软件和工具

蛋白质组定量数据处理需要多个步骤. 到目前为止, 已开发出许多开源的定量工具包, 如 MaxQuant[62], SILVER[63]和 Multi-Q[64]等. 它们有不同的适用范围, 但大部分只是简单地计算蛋白的丰度比. 随着候选生物标记物研究的需要, 近年来, 定量工具的开发愈加重视差异蛋白筛选的统计学方法和分析结果的质量控制部分. 为了使统计方法能更广泛地应用于蛋白质组学定量数据分析中, 并达到高质量分析数据的目的, 研究者针对基于质谱的定量结果开发出一系列程序软件, 以方便相关研究人员使用. 表 3 列出了几种有代表性的利用统计学原理筛选差异表达蛋白的工具.

4 结语

毫无疑问, 差异表达蛋白的筛选对于进一步开展生物标记物的发掘至关重要, 而统计学方法为这一问题的解决提供了新思路, 但同时也面临许多挑战. 针对基于质谱的定量蛋白质组学研究的 3 大困难, 本文主要分析总结了已提出方法及工具的特点, 并对其中有代表性的方法进行了简单评述.

在基于质谱的定量过程中, 肽段的信号峰通常会缺失, 数据信息不完整会影响蛋白丰度的准确性. 因此, 对蛋白表达水平进行统计推断前, 可以先进行过滤, 以选择合适的肽段数据集, 也可以选用参数估计方法对缺失值进行填充, 为进一步进行数据分析做准备.

当前, 不少蛋白质组学的研究通常仅进行单次或 2 次重复实验, 然而已提出的方法中不少都需要进行至少 3 次重复, 这使得大部分方法在实际情况中并不能真正起作用. 所以, 在数据分析前, 建议先确定样品的重复实验次数, 再选择统计方法, 如非参数统计检验方法, 或者是结合多种统计检验方法, 结果取交集, 互相弥补不足. 并且, 针对那些适用于重复实验次数较多的方法, 必须进行优化改进, 使它们更适合蛋白质组学的数据分析特点.

差异蛋白筛选的过程中会产生大量假阳性结果, 导致最终得到的差异蛋白较多, 候选生物标记物因此也不可靠, 这对进一步推测生物标记物来说带来一定的挑战. 针对此问题, 一方面, 可以使用投票的策略在一定程度上减少假阳性蛋白的数量; 另一方面, 由于统计方法的发展远不能满足基于质谱的定量差异蛋白的研究需要, 而实验结果的质量控制不可忽视, 严格控制出现的检验错误率就显得十分重要. 这就需要有统计背景的相关人员对多重假设检

表 3 用于筛选差异表达蛋白的工具

软件名称	简介	特点	使用方法	所属类别	参考文献
ReSASC	针对基于谱图计数的定量数据, 利用重新抽样算法, 开发出的一种筛选蛋白的工具	在实验结果中能提供更多有效信息. 但目前只能比较 2 状态间的差异, 且要求每个状态至少进行 3 次重复实验	此算法已被编译成程序在 Matlab(R2007b)中运行	基于经典统计检验策略	[65]
TFold	属于 PatternLab 的一个模块: 特征选择	可检测到低丰度的蛋白, 对实验的重复次数没有严格要求	http://pcarvalho.com/patternlab	其他策略	[66]
ROTs	给定一组数据, 程序就可以根据数据的特征, 选出合适的排序统计量, 通过随机置换样本标签计算最优统计量的 *FDR*	检验时无需数据的任何先验信息和重复实验, 且允许数据存在缺失值, 能够鉴定到传统分析没有检测到的几个肽段标记物	途径 1. 单机 R 程序 (http://www.math.utu.fi/short/rots.html) 途径 2. 与开源软件 Chipster 整合 (http://chipster.sourceforge.net/)	基于经典统计检验策略	[67]
PepC	综合 t 检验和 G 检验 2 种统计方法, 共同分析基于谱图计数的定量结果	可最大化鉴定的差异表达蛋白的数目, 且有效控制实验的假阳性率	运行环境 Java(http://sashimi.svn.sourceforge.net/viewvc/sashimi/trunk/trans_proteomic_pipeline/src/Quantitation/Pepc)	基于经典统计检验策略	[68]
DanteR	在 DanTE 的基础上改进, 包括以下模块: 归一化、假设检验、动态可视化和肽段/蛋白的归纳	可在肽段、蛋白水平上进行显著性分析, 使用者可以在软件中增加自己的算法	R 程序包(http://omics.pnl.gov/software/)	基于经典统计检验策略	[23]

验问题进行深入研究.

总之, 在以后的研究中, 针对基于质谱的定量蛋白质组学研究的 3 大挑战: 多缺失、少重复、实验结果不可靠, 必须同时考虑前 2 个问题, 优化改进现有策略, 重视数据分析中的多重假设检验问题, 进一步完善对蛋白定量结果的统计推断, 最后提出一种更适合差异蛋白分析的统计方法, 以选择更可靠的候选生物标记物.

参考文献

1 Werner T. Promoters can contribute to the elucidation of protein function. Trends Biotechnol, 2003, 21: 9–13
2 Geiger T, Wehner A, Schaab C, et al. Comparative proteomic analysis of eleven common cell lines reveals ubiquitous but varying expression of most proteins. Mol Cell Proteomics, 2012, 11: M111.014050
3 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: 1902–1909
4 张尧庭. 统计中的三大学派. 统计教育, 1995, 1: 35–39
5 Student. On the error of counting with a haemacytometer. Biometrika, 1907, 5: 351–360
6 Roxas BA, Li Q. Significance analysis of microarray for relative quantitation of LC/MS data in proteomics. BMC Bioinformatics, 2008, 9: 187
7 Andreev V P, Li L, Rejtar T, et al. New algorithm for 15N/14N quantitation with LC-ESI-MS using an LTQ-FT mass spectrometer. J Proteome Res, 2006, 5: 2039–2045
8 Wu C C, MacCoss M J, Howell K E, et al. Metabolic labeling of mammalian organisms with stable isotopes for quantitative proteomic analysis. Anal Chem, 2004, 76: 4951–4959
9 单文娟, 童春发, 施季森. 基因芯片筛选差异表达基因方法比较. 遗传, 2008, 30: 1640–1646
10 Tusher V G, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci USA, 2001, 98: 5116–5121
11 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: 1945–1951
12 Allet N, Barrillat N, Baussant T, et al. *In vitro* and *in silico* processes to identify differentially expressed proteins. Proteomics, 2004, 4: 2333–2351
13 Cho H, Smalley D M, Theodorescu D, et al. Statistical identification of differentially labeled peptides from liquid chromatography tandem mass spectrometry. Proteomics, 2007, 7: 3681–3692
14 Clough T, Thaminy S, Ragg S, et al. Statistical protein quantification and significance analysis in label-free LC-MS experiments with complex designs. BMC Bioinformatics, 2012, 13: S6
15 Cameron A C, Trivedi P K. Regression Analysis of Count Data. Cambridge: Cambridge University Press, 1998
16 Leitch M C, Mitra I, Sadygov R G. Generalized linear and mixed models for label-free shotgun proteomics. Stat Interface, 2012, 5: 89–98
17 Li M, Gray W, Zhang H, et al. Comparative shotgun proteomics using spectral count data and quasi-likelihood modeling. J Proteome Res, 2010, 9: 4295–4305
18 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: 363–369
19 Pham T V, Jimenez C R. An accurate paired sample test for count data. Bioinformatics, 2012, 28: I596–I602
20 Fisher R. Statistical Methods for Research Workers. Edinburgh: Oliver and Boyd, 1934
21 Sokal R R, Rohlf F J. Biometry: the Principles and Practice of Statistics in Biological Research. 3rd ed. New York: W. H. Freeman and Company, 1995
22 Audic S, Claverie J M. The significance of digital gene expression profiles. Genome Res, 1997, 7: 986–995
23 Taverner T, Karpievitch Y V, Polpitiya A D, et al. Danter: an extensible R-based tool for quantitative analysis of -omics data. Bioinformatics, 2012, 28: 2404–2406
24 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: 2909–2918
25 Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Roy Stat Soc B, 1995, 57: 289–300
26 Cooper B, Feng J, Garrett W M. Relative, label-free protein quantitation: spectral counting error statistics from nine replicate mudpit samples. J Am Soc Mass Spectrom, 2010, 21: 1534–1546
27 Mann M. Comparative analysis to guide quality improvements in proteomics. Nat Methods, 2009, 6: 717–719

28 Troyanskaya O G, Garber M E, Brown P O, et al. Nonparametric methods for identifying differentially expressed genes in microarray data. Bioinformatics, 2002, 18: 1454–1461
29 Efron B. Microarrays, empirical bayes and the two-groups model. Stat Sci, 2008, 23: 1–22
30 Margolin A A, Ong S E, Schenone M, et al. Empirical bayes analysis of quantitative proteomics experiments. PLoS One, 2009, 4: e7454
31 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: 376–386
32 Koopmans F, Cornelisse L N, Heskes T, et al. Empirical bayesian random censoring threshold model improves detection of differentially abundant proteins. J Proteome Res, 2014, 13: 3871–3880
33 Troyanskaya O, Cantor M, Sherlock G, et al. Missing value estimation methods for DNA microarrays. Bioinformatics, 2001, 17: 520–525
34 Candes E J, Plan Y. Matrix completion with noise. Proc IEEE, 2010, 98: 925–936
35 Choi H, Fermin D, Nesvizhskii A I. Significance analysis of spectral count data in label-free shotgun proteomics. Mol Cell Proteomics, 2008, 7: 2373–2385
36 Booth J G, Eilertson K E, Olinares P D, et al. A bayesian mixture model for comparative spectral count data in shotgun proteomics. Mol Cell Proteomics, 2011, 10: M110.007203
37 Serang O, Paulo J, Steen H, et al. A non-parametric cutout index for robust evaluation of identified proteins. Mol Cell Proteomics, 2013, 12: 807–812
38 Serang O, Cansizoglu A E, Kall L, et al. Nonparametric bayesian evaluation of differential protein quantification. J Proteome Res, 2013, 12: 4556–4565
39 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
40 Pavelka N, Fournier M L, Swanson S K, et al. Statistical similarities between transcriptomics and quantitative shotgun proteomics data. Mol Cell Proteomics, 2008, 7: 631–644
41 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
42 Webb-Robertson B J, McCue L A, Waters K M, et al. Combined statistical analyses of peptide intensities and peptide occurrences improves identification of significant peptides from ms-based proteomics data. J Proteome Res, 2010, 9: 5748–5756
43 孙薇, 贺福初. 差异蛋白质组学研究技术新进展. 化学通报, 2005, 68: 401–407
44 Karpievitch Y, Stanley J, Taverner T, et al. A statistical framework for protein quantitation in bottom-up MS-based proteomics. Bioinformatics, 2009, 25: 2028–2034
45 Wang X, Anderson G A, Smith R D, et al. A hybrid approach to protein differential expression in mass spectrometry-based proteomics. Bioinformatics, 2012, 28: 1586–1591
46 Xu J, Wang L, Li J. A biological network module-based model for the analysis of differential expression in shotgun proteomics. J Proteome Res, 2014, doi: 10.1021/pr5007203
47 Pawitan Y, Murthy K R, Michiels S, et al. Bias in the estimation of false discovery rate in microarray studies. Bioinformatics, 2005, 21: 3865–3872
48 Shaffer J P. Multiple hypothesis testing. Annu Rev Psychol, 1995, 46: 561–584
49 Storey J D. The positive false discovery rate: a bayesian interpretation and the q-value. Ann Stat, 2003, 31: 2013–2035
50 Dudoit S, Yang Y H, Callow M J, et al. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Stat Sinica, 2002, 12: 111–140
51 Fu Y, Qian X. Transferred subgroup false discovery rate for rare post-translational modifications detected by mass spectrometry. Mol Cell Proteomics, 2013, 13: 1359–1368
52 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: 207–214
53 Tan Y D, Xu H. A general method for accurate estimation of false discovery rates in identification of differentially expressed genes. Bioinformatics, 2014, 30: 2018–2025
54 Farcomeni A. A review of modern multiple hypothesis testing, with particular attention to the false discovery proportion. Stat Methods Med Res, 2008, 17: 347–388
55 Efron B. Large-scale simultaneous hypothesis testing. J Am Stat Assoc, 2004, 99: 96–104
56 Storey J D, Tibshirani R. Statistical significance for genomewide studies. Proc Natl Acad Sci USA, 2003, 100: 9440–9445
57 Bei Y, Hong P. A novel approach to minimize false discovery rate in genome-wide data analysis. BMC Syst Biol, 2013, 7: S1
58 Noble W S. How does multiple testing correction work? Nat Biotechnol, 2009, 27: 1135–1137

59 Sham P C, Purcell S M. Statistical power and significance testing in large-scale genetic studies. Nat Rev Genet, 2014, 15: 335–346
60 Krzywinski M, Altman N. Points of significance: importance of being uncertain. Nat Methods, 2013, 10: 1041–1042
61 Carvalho P C, Fischer J S, Chen EI, et al. Patternlab for proteomics: a tool for differential shotgun proteomics. BMC Bioinformatics, 2008, 9: 316
62 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: 1367–1372
63 Chang C, Zhang J, Han M, et al. Silver: an efficient tool for stable isotope labeling LC-MS data quantitative analysis with quality control methods. Bioinformatics, 2014, 30: 586–587
64 Lin W T, Hung W N, Yian Y H, et al. Multi-*q*: a fully automated tool for multiplexed protein quantitation. J Proteome Res, 2006, 5: 2328–2338
65 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: 1212–1222
66 Carvalho P C, Yates J R 3rd, Barbosa V C. Improving the TFold test for differential shotgun proteomics. Bioinformatics, 2012, 28: 1652–1654
67 Elo L L, Hiissa J, Tuimala J, et al. Optimized detection of differential expression in global profiling experiments: case studies in clinical transcriptomic and quantitative proteomic datasets. Brief Bioinform, 2009, 10: 547–555
68 Heinecke N, Pratt B, Vaisar T, et al. PepC: proteomics software for identifying differentially expressed proteins based on spectral counting. Bioinformatics, 2010, 26: 1574–1575

Statistical Strategies for Selection of Differentially Expressed Proteins Based on Mass Spectrometry Technology

WANG JinXia[1], CHANG Cheng[2], MA Jie[2], WU SongFeng[2], ZHUANG JuJuan[1] & ZHU YunPing[2]

1 Mathematics Department, Dalian Maritime University, Dalian 116026, China;
2 State Key Laboratory of Proteomics, Beijing Proteome Research Center, Beijing Institute of Radiation Medicine, Beijing 102206, China

With rapid development of mass spectrometry, proteomics has become a popular research field following genomics and transcriptomics. Searching for reliable differentially expressed proteins will be crucial to biomarker discovery. Therefore, how to find out differentially expressed proteins accurately and sensitively has become one of the most important subjects in quantitative proteomics research. At present, there are a few research strategies focusing on this issue, but these methods have different applicable scopes and limitations. In general, the statistical strategies for selection of differentially expressed proteins based on mass spectrometry technology have three categories: the statistical strategies based on classical school, the statistical test strategies based on Bayesian school and the others. These methods differ in the application scopes, features and disadvantages. In addition, some false positive results will be generated during the process of selection. To improve the reliability of the results, new methods are needed with the development of quantitative proteomics.

mass spectrometry, proteomics, differentially expressed protein, statistics theory, multiple hypothesis testing

doi: 10.1360/N052014-00197