中国科学: 生命科学 2018 年 第 48 卷 第 3 期: 278 ~ 286
SCIENTIA SINICA Vitae lifecn.scichina.com
《中国科学》杂志社 SCIENCE CHINA PRESS

评 述

小蛋白质鉴定研究进展

何崔同[1,2], 张瑶[2,3*], 徐平[1,2*]

1. 安徽医科大学研究生学院, 合肥 230032;
2. 军事科学院军事医学研究院生命组学研究所, 蛋白质组学国家重点实验室, 国家蛋白质科学中心(北京), 北京蛋白质组研究中心, 北京 102206;
3. 中山大学生命科学学院, 广州 510275
* 联系人, E-mail: zhangyaowsw@163.com; xupingghy@gmail.com

收稿日期: 2017-10-23; 接受日期: 2017-11-17; 网络发表日期: 2018-01-29
精准医学研究重点专项(批准号: SQ2017YFSF090210)、国家自然科学基金(批准号: 31400698)和北京市科技领军人才(批准号: Z161100004916024)资助

摘要 小蛋白质是指可读框编码长度小于100个氨基酸的多肽. 近年来研究发现, 小蛋白质参与了细胞信号转导、代谢和生长等重要的生物学过程. 然而, 由于在基因组注释和生化检测上存在技术挑战, 小蛋白质研究进展较为缓慢. 小蛋白质高效鉴定技术的发展是功能研究的前提, 也是完善基因组注释的必然要求. 本文系统综述了小蛋白质鉴定的难点、原因和解决方法.

关键词 蛋白质组学, 小蛋白质, 鉴定

尽管成千上万蛋白质的结构与功能已被系统地研究, 但由小可读框(short open reading frame, sORF)编码的小蛋白质(small proteins, SPs)研究仍被极大地忽视了. 酵母基因组测序后, 研究者以最小100个密码子为限来注释其潜在ORF[1]. 这是出于提高注释可信度的考虑, 若将理论2~99个密码子的ORF都注释将会产生额外的260000个ORF[2]. 于是, 这种编码基因100个密码子的最小限制也被广泛应用到其他真核生物基因组注释中[3].

由于细菌的基因组较小, 其最小注释标准稍小些, 但仍然导致了大量小蛋白质编码基因的漏注释[3]. 随着广泛的单基因敲除验证[4]、转录组结果的补充[5]以及蛋白质组学的相关研究[6]等, 包括sORF在内的一些漏注释基因逐渐被发现、验证和重注释. 近年来的许多研究也显示, 小蛋白质具有重要的生物学功能, 如小蛋白质可以通过进入其他蛋白质的别构位点调节蛋白互作和酶活性[7], 作为细胞因子参与信号转导和细胞间通讯[8], 以及作为基因表达的转录后调节器[9].

小蛋白质的鉴定是功能研究的前提, 高效小蛋白质鉴定技术的发展有助于发现更多漏注释sORF. 小蛋白质鉴定的策略主要有两种: (ⅰ) 基于荧光[10]和标签[11]的方法; (ⅱ) 基于质谱的蛋白质组学方法. 前者可以直观观察小蛋白质的细胞定位和表达情况, 但难以实现高通量鉴定, 且一个潜在的问题是标签的引入可能会影响小蛋白质的结构和功能[12]; 而蛋白质谱可以在组学水平高通量准确鉴定蛋白质. 然而, 据UniProt数据库记录, 目前几乎所有物种的大多数已注释小蛋白质尚缺少蛋白水平存在的证据和相关功能研

引用格式: 何崔同, 张瑶, 徐平. 小蛋白质鉴定研究进展. 中国科学: 生命科学, 2018, 48: 278–286
He C T, Zhang Y, Xu P. Advances in small protein identification (in Chinese). Sci Sin Vitae, 2018, 48: 278–286, doi: 10.1360/N052017-00245

究，小蛋白质的高效鉴定一直以来是研究者无法回避的难题. 本文根据本研究组前期的研究，从真核生物中小蛋白质的表达、稳定性、提取、分离富集、消化和数据库搜索层面综述小蛋白质在质谱鉴定中存在的难点、原因和可能的解决方法.

1　小蛋白质的表达

小蛋白质的表达是小蛋白质鉴定的前提. 其难点是很多小蛋白质的表达可能是非常规条件依赖的. 因此，为了促进小蛋白质的表达，需要了解小蛋白质的翻译机制以及翻译的调控机制，在此基础上去寻找促进小蛋白质表达的合适条件.

1.1　小蛋白质的翻译机制

sORF通过它们的大小与其他的ORF区别开，但并非所有的sORF都具有翻译潜能. 具有翻译潜能的sORF已被发现可以位于已知ORF的5′UTR或3′UTR, 也可位于已知ORF内部或与已知ORF重叠，另外也可位于之前被视为非编码的RNA上，包括一些长的非编码RNA、基因间转录本和反义转录本[13].

位于mRNA 5′UTR区的sORF也被称为上游ORF (upstream ORF, uORF). uORF翻译的机制包括重启动机制和渗漏扫描机制[13]. 重启动机制是指核糖体40S小亚基从mRNA的5′末端开始扫描，当遇到uORF的起始密码子便与60S亚基生成80S起始复合物，从而开始uORF的翻译. 完成翻译后，60S亚基解离，40S亚基仍保持与mRNA的结合并继续向下游扫描，当遇到下游ORF的起始密码子时再次启动翻译. 渗漏扫描机制是指核糖体40S小亚基既可以识别uORF的起始密码子并启动翻译, 也可以忽略uORF的起始密码子而在下游起始密码子处启动翻译. 在uORF的起始密码子处起始翻译的核糖体的数目由该起始密码子周围的序列决定.

真核生物以外其他类型sORF的翻译机制研究尚少. 例如, 位于3′UTR区的sORF可以在转录本上游被剪切后翻译，这可能是由于暴露了3′尾随序列的核糖体结合位点所致[14].

1.2　小蛋白质翻译的调控

目前在小蛋白质翻译的调控机制上的研究较少，以uORF为例介绍小蛋白质翻译的调控方法和机制.

(1) 稳态失衡诱导uORF发挥抑制元件功能. 在高水平精氨酸存在的条件下，酵母CPA1转录本的uORF表达精氨酸衰减肽AAP，精氨酸与仍与核糖体结合的25个氨基酸的AAP相互作用，阻止核糖体的迁移，从而抑制了下游CPA1的表达[15,16]. 在高水平蔗糖存在条件下的*BZIP11*转录本[17]、高水平多胺存在条件下的*SAMDC*转录本[18]、高水平磷酸胆碱存在条件下的*XPL1*转录本[19], 其uORF均可表达小肽. 这些小分子和初生的小肽相互作用均可使得核糖体停滞在uORF, 从而阻碍核糖体进程和下游ORF的翻译. 这实际上反映了uORF作为基因表达的抑制性元件功能, 已在许多基因或基因组水平的研究中得到验证[9].

深入研究发现，酵母*CPA1*基因编码了参与精氨酸生物合成的氨甲酰磷酸合成酶的一个亚基[16]，拟南芥(*Arabidopsis thaliana*) *SAMDC*基因编码了参与多胺生物合成的S-腺苷甲硫氨酸脱羧酶[20]，拟南芥*XPL1*基因编码了参与磷酸胆碱生物合成的磷酸乙醇胺*N*-甲基转移酶1[21]，也即促使uORF表达的高水平环境信号与它们下游ORF表达蛋白的功能效应是一致的. 因此在同一转录本中下游ORF编码蛋白的功能效应被过度放大的情况下，为维持机体稳态，会激发uORF表达调节性小肽, 并通过与小分子相互作用来干扰核糖体进程, 从而抑制下游ORF的表达.

这些现象可能说明uORF的表达可能是非常规条件依赖的. 利用uORF和下游ORF的这种负反馈调控，人为顺势增强下游ORF编码蛋白的终效应，可能促进uORF的表达.

(2) 可变转录起始位点的使用和可变剪接. 可变转录起始位点的使用和可变剪接可以将uORF选择性地包含进转录本或从转录本排除, 这是对uORF调控的最直接的一种形式[9]. 在成肌细胞分化为肌管的过程中存在着不同的转录起始位点使用情况. 例如，对于Cryab的转录本, 在成肌细胞中检测到两种转录起始位点, 而在肌管中则只检测到其中一种转录起始位点. 前者的其中一个转录起始位点位于5′UTR，包含进一个uORF，而后者的转录起始位点则将uORF排除出去. uORF的存在抑制了Cryab的表达，uORF的排除则促进了Cryab的表达. Cryab则进一步通过促进MyoD的表达来调节了肌分化[22]. 因此, 这种不同转录起始位点的使用以包含或排除uORF符合了成肌细胞分化的需要.

促血小板生成素(thrombopoietin, *THPO*)基因的一

个可变剪接体位点的突变形成缺少uORF的*THPO*转录本. 这种缺失uORF的转录本比正常的转录本更高效地翻译THPO, 产生过量的THPO蛋白, 从而引起血小板增多症[23].

这些例子都是uORF作为抑制性元件, 决定了可变转录起始位点的选择性使用或使可变剪接作为uORF发挥抑制性元件的手段. 因此, 可能通过控制环境信号来影响uORF抑制性元件功能的发挥, 从而间接调控可变转录起始位点的使用和可变剪接.

(3) 翻译相关因子的活性调节. 一些RNA结合因子或翻译机器相关蛋白可以通过调控重启动或渗漏扫描来影响uORF表达[24]. 一个促进上游AUG起始子选择的因子是DExH盒解旋酶DHX29, 它可以通过和翻译起始因子eIF1A作用来改变翻译起始复合体的构象, 使之稳定而减少渗漏扫描[25]. 此外, 改变翻译起始因子eIF2的活性和浓度也可影响核糖体的重启动翻译. 在多种压力条件下, 真核细胞通过eIF2的磷酸化介导的失活迅速减少蛋白质合成, 减少细胞能量的消耗, 度过外界压力刺激难关. 如人细胞系在亚砷酸钠处理时, eIF2可以快速被磷酸化. 核糖体图谱分析表明, 虽然绝大部分基因的翻译被抑制了5.4倍, 但某些转录本仍然表现出对抑制的抵抗. 这些转录本都有至少一个被高效翻译的uORF, 且抑制了正常生理条件下主体ORF的翻译. uORF在对抑制的抵抗中发挥重要功能[26].

还有一些RNA结合因子可以影响特定uORF的翻译. 如RNA结合蛋白SXL可与msl-2转录本中uORF的下游结合, 通过促进uORF的翻译起始而减少了渗漏扫描, 使得uORF对下游翻译的抑制效应提高了9倍[27].

2 小蛋白质的稳定性

小蛋白质可能不稳定, 半衰期短, 也增加了鉴定的难度. Johnstone等人[28]在人、小鼠(*Mus musculus*)和斑马鱼(*Danio rerio*)中发现uORF转录本的稳定性普遍较低. 无义介导的mRNA降解(nonsense-mediated mRNA decay, NMD)是广泛存在于真核生物细胞中的一种mRNA质量监控机制, 该机制通过识别和降解含有提前终止密码子的转录产物进而防止有潜在毒性的截短蛋白的产生[29]. 在酵母和哺乳动物细胞中有许多含uORF的转录本是NMD的底物[30,31]. 与正常mRNA降解途径不同, NMD不依赖于Poly(A)尾巴的水解缩短而直接进行5′端脱帽和5′→3′方向的水解[32]. 识别含提前终止密码子的mRNA的关键是监视复合体在该mRNA上的形成. 其中参与监视复合体形成的核心因子是系列UPF蛋白, 包括UPF1, UPF2和UPF3[33]. 研究发现, 在野生型酵母细胞中*CPA1*转录本的半衰期是3 min, 而在*UPF1*敲除的菌株中则为18 min[34], 证明*CPA1*转录本是NMD的底物. 转录本中提前终止密码子的出现并不足以触发它本身的降解, *CPA1*转录本的NMD的发生依赖于uORF的翻译, uORF表达的精氨酸衰减肽与核糖体作用引起其在uORF终止密码子的停滞诱导了NMD的发生[15]. 进一步研究发现, 将精氨酸衰减肽的第13位天冬氨酸突变为天冬酰胺, 可减弱延滞核糖体的能力, 并导致NMD的相对失活; 而通过优化uORF前的翻译起始序列, 提升延滞核糖体的能力, 可促进NMD的激活. 这表明通过调控提前终止密码子上的核糖体水平可以控制NMD[15].

在酵母中分别构建NMD相关因子*UPF1*, *UPF2*, *UPF3*, *DCP1*和*XRN1*的单敲除菌株, 并进行全基因组范围内的表达谱分析, 发现除了*XRN1*的敲除只引起一小部分RNA的上调外, 其他均可引起整体RNA的积累. 对上调的核心转录本进行分析, 发现含uORF的转录本, 如*CPA1*在*UPF2*敲除的情况下上调近7倍, *EST1*在*UPF1*敲除的情况下上调14倍, *PET130*在*XRN1*敲除的情况下上调近10倍, *THI20*在*DCP1*敲除的情况下上调近9倍[30]. 这说明可能通过抑制NMD相关因子来延长含uORF的转录本的半衰期, 促进uORF的表达.

小蛋白质本身的低稳定性也是限制小蛋白质鉴定的重要原因. 但由于技术限制, 这方面的研究至今尚少. 在基因工程表达小蛋白质的研究中, 由于其酶降解性高、半衰期短等特点, 常以融合蛋白的形式表达[35]. 胞内蛋白质降解主要通过自噬溶酶体和泛素蛋白酶体系统两种途径进行[36]. 考虑到小蛋白质大小的特殊性以及小蛋白质常作为调控分子或在特定条件下起作用, 因此猜测小蛋白质可能存在着特殊的降解机制, 因此其降解机制亟待研究.

3 小蛋白质的提取

小蛋白质提取包括细胞破碎和蛋白质溶出两个方面. 影响小蛋白质提取效率的因素主要包括细胞的破碎效率、小蛋白质的溶解性、单次提取蛋白的释放

率、小蛋白质的亚细胞定位以及小蛋白质在提取过程中的稳定性等5个方面.

3.1　细胞破碎

细胞破碎的方法有很多，如机械破碎、超声破碎、表面活性剂裂解、高压破碎等[37]. 对于具有厚细胞壁的酵母，常采用玻璃珠进行机械破碎. 如裂解液中含脱氧胆酸钠, 超声破碎的方法可通过切碎核酸链, 减少长链核酸的黏稠, 提升蛋白质提取效率.

3.2　小蛋白质的溶解性

天然状态的蛋白质由于二硫键、氢键、离子和疏水相互作用的存在, 蛋白质易聚集, 溶解度低, 需要特定条件使蛋白质以独立的多肽形式分散进溶液，促进其溶解[38]. 离液剂(如尿素和硫脲)和表面活性剂(如CHAPS和Triton X-100)常用于蛋白质的增溶[37]. 离液剂通过破坏氢键和亲水相互作用使蛋白质解折叠，使所有可离子化基团暴露于溶液，促进蛋白质溶解. 尿素是一种中性离液剂，5~9 mol/L的高浓度尿素可以有效地破坏蛋白质的二级结构. 如在尿素溶液中加入适量硫脲，可以显著促进蛋白质的溶解. 不过含这类离液剂的样品在处理时温度不可高于37℃，否则尿素和硫脲会水解为氰酸酯和硫氰酸，修饰蛋白质本身[39]. 表面活性剂则通过其两亲性破坏疏水相互作用，从而显著促进疏水蛋白质如膜蛋白的溶解.

小蛋白质中疏水蛋白占了较大比例[3]，如根据酵母基因组数据库注释，酵母小蛋白质中疏水蛋白占比达到43%. 因此, 高效的增溶方式对小蛋白质的提取尤为重要.

3.3　单次提取的蛋白释放率

尽管提取蛋白质的方法很多，但在实验中发现单次提取的总蛋白质，其释放率总是偏低. 如使用不同的溶剂进行多次萃取, 可以弥补这一缺陷[40].

3.4　小蛋白质的亚细胞定位对蛋白质提取的影响

根据UniProt对蛋白质的亚细胞定位注释信息发现酵母的小蛋白质中, 膜蛋白占了60%, 细胞质蛋白占了18%，核蛋白占了10%. 在人的小蛋白质中，分泌蛋白占了36%, 膜蛋白占了35%, 细胞质蛋白占了14%. 这表明小蛋白质存在着亚细胞定位的偏好性，且不同物种中的定位偏好性有所不同.

通常蛋白质的提取都是以完整的细胞为研究对象. 由于胞质蛋白易于提取, 使得非细胞质定位小蛋白质易于漏检. 如辅以针对亚细胞定位特征的小蛋白质提取策略, 有助于促进小蛋白质的提取[6].

由于疏水作用, 一些小蛋白质可能与脂质结合. 这种结合态可能会干扰小蛋白质的消化. 因此提取后的进一步除杂质，如甲醇沉淀蛋白，氯仿抽提脂质等都会有助于小蛋白质的鉴定[41].

3.5　小蛋白质在提取过程中的稳定性的影响

酵母中有170多种蛋白酶及其同源物[42]. 这些酶如在样品制备过程中不被抑制，会引起总细胞蛋白质不受控的非特异性降解. 抑制蛋白酶活性的方法包括快速小规模的提取、水煮、微波辐射、有机溶剂变性及加入蛋白酶抑制剂等[37]. 使用70℃的水处理细胞20 min可以在不引起非特异降解的情况下较好抑制蛋白酶活性, 而用70℃的10 mmol/L HCl处理细胞20 min则会引起蛋白序列中天冬氨酸和脯氨酸间的肽键断裂[43]. 因此在酸处理提取小蛋白质时, 需要避免热酸效应可能导致的非特异降解.

4　小蛋白质的分离和富集

为了避免高丰度大蛋白质的信号掩盖，需要将小蛋白质从复杂提取物中分离出来. 但小蛋白质在分离时易被丢失，需要特殊的分离方法. 分离方法可基于分子大小、溶解度、带电性质和配体特异性来展开.

基于分子大小的分离方法有许多，如分子排阻色谱法可以分离特定质量范围的蛋白质. 这种方法的缺点是耗时长[44]. 分子截留膜过滤则是最简单的小蛋白质分离方法，但因为规格的限制，并不能做到任意指定质量范围的蛋白质的分离，而且会存在一定的损失[45].

基于溶解度的分离方法主要是选择性沉淀. 酸[46]、不同的有机溶剂[47]或这些沉淀剂的组合[48]常用于实现此目的. 然而, 蛋白沉淀不像膜过滤一样理论上完全去除特定分子量范围的蛋白质，有些小蛋白质可能发生聚集, 并在沉淀物中损失[49].

基于带电性质的分离方法有电泳和离子交换层析. 传统的二维凝胶电泳主要适用于等电点4~10和分

子量10~200 kD的中等大小的蛋白质的分离[50]. Tricine凝胶电泳是分离小蛋白质的常用方法[51]. 但由于凝胶电泳的载样量的限制, 需要对小蛋白质进行分离富集. 微量级制备电泳技术是基于电洗脱的原理从凝胶中回收蛋白质[52].

在基于配体的研究策略方面, 已有纯化含有特定氨基酸残基如半胱氨酸[53]、色氨酸[54]或甲硫氨酸[55]的多肽的方法. 一些方法基于磷酸肽与二氧化钛的相互作用差异来分离磷酸肽[56], 以三磷酸腺苷为金属载体对磷酸肽的分离有着高敏感性和选择性[57]. 还有一些方法基于亲水作用液相色谱, 采用不同固定相如微晶纤维素[58]、两性离子材料[59]等来分离糖肽.

不同小蛋白质的理化性质如大小、疏水性和净电荷等均不同, 它们理化性质的多样性通过翻译后修饰进一步增加. 要想一次性从复杂提取物中全面分离小蛋白质是具有挑战性的, 多种分离方法的使用可能有助于改善该问题.

5 小蛋白质的消化

小蛋白质由于疏水性和大小的限制, 存在切割效率低的风险, 加上产生质谱易鉴定的7~25个氨基酸残基的肽段少, 鉴定难. 提高消化效率是其中的关键.

5.1 组合策略提高消化效率

根据切割主体的类型可以将消化方式分为基于特异蛋白酶的酶解和基于化学试剂的化学切割两类. 用多种酶组合或平行地进行酶切的策略可以提高蛋白序列覆盖度[60]. 如传统胰酶消化下的C端肽段常缺少碱性氨基酸, 离子化效率低, 二级谱图质量差, 不易被鉴定. 胰酶的镜像酶可在精氨酸或赖氨酸的N端切割, 有助于提高C端覆盖度[61]. 化学切割可用于一些蛋白酶难消化蛋白质样品的制备. 如采用CNBr化学切割结合Trypsin消化可以有效促进膜蛋白的鉴定[62].

5.2 消化条件的优化

通过优化消化条件可以改善疏水蛋白的消化效率. 如细菌视紫红质是有7个跨膜结构域的膜蛋白, 在0.01% SDS或10%甲醇组合0.01% RapiGest条件下使用Trypsin消化均可达到40%的蛋白序列覆盖度, 而在常规消化条件下难于鉴定[63]. 2%~5%的脱氧胆酸钠在促进蛋白溶解的同时也可以提高胰酶对疏水蛋白的消化效率[64].

5.3 胶内或溶液消化的选择

目前蛋白质组学中蛋白质的消化常在胶内或溶液中进行. 胶内消化的优势在于可以排除影响消化的杂质因素, 但是可能影响肽段的回收且不易自动化. 为了避免小蛋白质在样品处理过程中可能从胶孔中漏失, 可以选用溶液消化的方法. 溶液消化便于自动化且处理步骤简单, 但蛋白质组可能存在溶解不完全的情况, 且消化易受杂质干扰. 实际上, 像去污剂, 即便很小的浓度, 可以阻碍酶的消化. 这些去污剂由于高离子化效率而占据较多的质谱离子注入空间, 干扰质谱鉴定. 为了避免杂质如SDS对溶液消化的影响, 一个以置换溶剂为思路的过滤管协助的样品准备方法(FASP酶切)已被广泛用于溶液消化中, 有效提高了蛋白质组覆盖度[65]. 当然, 这里应该使用截留分子量小的过滤管, 否则会引起小蛋白质的损失.

5.4 “自顶向下”的策略

目前广泛采用的通过酶切产生的肽段倒推蛋白质的方法称为“自底向上”的策略, 而对于一些在消化下不产生或难产生合适唯一肽段的小蛋白质, 可以采用不酶切、直接以完整蛋白质为分析对象的“自顶向下”的鉴定策略[66]. “自顶向下”的策略依赖于完整蛋白质的高效分离技术, 可获得精确分子量和丰富二级碎裂谱信息的质谱技术以及相关生物信息技术. 但目前这种策略仍存在一定的技术难题, 使用不多.

6 数据库搜索

目前的蛋白质鉴定主要采用依赖已注释蛋白库的数据库搜索的方法. 数据库中不存在的蛋白质不能被鉴定. 因此, 小蛋白质鉴定在数据库搜索层面存在的难点是漏注释基因编码小蛋白质的鉴定. 完整数据库的构建和质控机制的完善是漏注释基因编码小蛋白质鉴定的两个要素.

6.1 完整数据库的构建

目前基因预测方法主要有两大类: (ⅰ) 基于同源性的方法, 以检索序列与已知基因序列最大的匹配为

基础; (ii) 基于从头算的方法, 以给定的序列本身来进行预测, 主要取决于人们对已知基因结构特征的认识[67]. 有研究用这两类方法分别比较了预测酵母中小于100个密码子的sORF和100~150个密码子的较大基因的准确性, 发现小基因预测的准确性相对大基因要差很多. 这主要是由于可用于同源性搜索和训练从头算技术的经实验验证的sORF的数据集过小所致. sORF预测准确性的提高就迫切地依赖于有关sORF的实验和算法层面的研究[68]. 为了保证数据库的完整性, 目前对于原核生物比较常用的方法是利用基因组数据通过六可读框翻译来构建新的蛋白库[69]. 对于真核生物因为可变剪接带来的复杂性, 六可读框翻译无法覆盖跨越剪接位点的肽段, 所以发展出直接枚举的方法[70]、基于序列标签的方法[71]和利用转录组数据[72]构建蛋白库等. 此外, 基于N端蛋白质组学技术和核糖体图谱技术在小鼠和人当中进行可变翻译起始位点的研究发现小蛋白质存在普遍的非经典翻译起始子的使用[73]. 因此, 对于构建六可读框翻译数据库, 为了防止理论酶切对N端肽段的遗漏以及小蛋白质非经典翻译起始位点使用的不确定性, 需要额外添加不同起始密码子起始的N端肽段.

6.2　完善质控

在相同FDR过滤标准下, 大数据库的引入会严重影响鉴定结果的灵敏度[74]并导致新肽段的高假阳性率[75]. Brosch等人[76]采用后验概率阈值(PEP<0.01)控制单个肽段的错误率. Branca等人[77]利用高分辨率等电聚焦电泳对肽段进行预分离, 将六可读框翻译库理论酶切肽段按照预测的等电点分成相应区间建立小库的方法来减小数据库规模, 并采取对新肽段单独过滤、统计类别FDR的策略来控制FDR在合理的范围内. Fu和Qian[78]的一项类别FDR工作从形式上给出了影响不同肽段FDR的因素. 进一步地, 本研究组和中国科学院计算技术研究所贺思敏研究员领导的研究组[79]合作研究了新肽段类别FDR与数据库规模以及注释完整性的定量关系, 发现基因组注释完整性比例是制约新肽段类别FDR准确性的主导因素, 在定量水平建立了采用分开过滤计算新肽段类别FDR的质控机制. 这些方法改善了新编码蛋白质鉴定时FDR失真的影响, 降低了假阳性.

7　小结与展望

与大蛋白质的充分研究相反, 小蛋白质的生物学研究仍是待开垦的领域. 尽管越来越多的具有重要生物学功能的小蛋白质被鉴定和发现, 大多数已被注释小蛋白质仍缺少蛋白水平存在的证据, 也无法开展深入的功能研究. 这主要是由于小蛋白质的鉴定是当前蛋白质组学领域的一个难点. 随着对小蛋白性质认识的不断深化、样品制备技术的改进以及质谱技术的发展, 针对小蛋白质开发出更加高效甚至全新的鉴定策略, 必会实现小蛋白质的深度覆盖. 这将为深入探究这些小蛋白质在生命活动中的功能奠定坚实的技术基础.

参考文献

1 Harrison P M, Kumar A, Lang N, et al. A question of size: the eukaryotic proteome and the problems in defining it. Nucleic Acids Res, 2002, 30: 1083–1090

2 Basrai M A, Hieter P, Boeke J D. Small open reading frames: beautiful needles in the haystack. Genome Res, 1997, 7: 768–771

3 Storz G, Wolf Y I, Ramamurthi K S. Small proteins can no longer be ignored. Annu Rev Biochem, 2014, 83: 753–777

4 Giaever G, Chu A M, Ni L, et al. Functional profiling of the *Saccharomyces cerevisiae* genome. Nature, 2002, 418: 387–391

5 Nagalakshmi U, Wang Z, Waern K, et al. The transcriptional landscape of the yeast genome defined by RNA sequencing. Science, 2008, 320: 1344–1349

6 Oshiro G, Wodicka L M, Washburn M P, et al. Parallel identification of new genes in *Saccharomyces cerevisiae*. Genome Res, 2002, 12: 1210–1220

7 Zanet J, Benrabah E, Li T, et al. Pri sORF peptides induce selective proteasome-mediated protein processing. Science, 2015, 349: 1356–1358

8 Costa L M, Marshall E, Tesfaye M, et al. Central cell-derived peptides regulate early embryo patterning in flowering plants. Science, 2014, 344: 168–172

9 Cabrera-Quio L E, Herberg S, Pauli A. Decoding sORF translation—from small proteins to gene regulation. RNA Biol, 2016, 13: 1051–1059

10 Huh W K, Falvo J V, Gerke L C, et al. Global analysis of protein localization in budding yeast. Nature, 2003, 425: 686–691

11 Ghaemmaghami S, Huh W K, Bower K, et al. Global analysis of protein expression in yeast. Nature, 2003, 425: 737–741

12 Sun Y H, de Jong M F, den Hartigh A B, et al. The small protein CydX is required for function of cytochrome bd oxidase in *Brucella abortus*. Front Cell Infect Microbiol, 2012, 2: 47

13 Andrews S J, Rothnagel J A. Emerging evidence for functional peptides encoded by short open reading frames. Nat Rev Genet, 2014, 15: 193–204

14 Mercer T R, Wilhelm D, Dinger M E, et al. Expression of distinct RNAs from 3′ untranslated regions. Nucleic Acids Res, 2011, 39: 2393–2403

15 Gaba A, Jacobson A, Sachs M S. Ribosome occupancy of the yeast CPA1 upstream open reading frame termination codon modulates nonsense-mediated mRNA decay. Mol Cell, 2005, 20: 449–460

16 Werner M, Feller A, Messenguy F, et al. The leader peptide of yeast gene *CPA1* is essential for the translational repression of its expression. Cell, 1987, 49: 805–813

17 Rahmani F, Hummel M, Schuurmans J, et al. Sucrose control of translation mediated by an upstream open reading frame-encoded peptide. Plant Physiol, 2009, 150: 1356–1367

18 Hanfrey C, Elliott K A, Franceschetti M, et al. A dual upstream open reading frame-based autoregulatory circuit controlling polyamine-responsive translation. J Biol Chem, 2005, 280: 39229–39237

19 Alatorre-Cobos F, Cruz-Ramírez A, Hayden C A, et al. Translational regulation of *Arabidopsis* XIPOTL1 is modulated by phosphocholine levels via the phylogenetically conserved upstream open reading frame 30. J Exp Bot, 2012, 63: 5203–5221

20 Ge C, Cui X, Wang Y, et al. BUD2, encoding an S-adenosylmethionine decarboxylase, is required for *Arabidopsis* growth and development. Cell Res, 2006, 16: 446–456

21 Cruz-Ramírez A, López-Bucio J, Ramírez-Pimentel G, et al. The xipotl mutant of *Arabidopsis* reveals a critical role for phospholipid metabolism in root system development and epidermal cell integrity. Plant Cell, 2004, 16: 2020–2034

22 de Klerk E, Fokkema I F A C, Thiadens K A M H, et al. Assessing the translational landscape of myogenic differentiation by ribosome profiling. Nucleic Acids Res, 2015, 43: 4408–4428

23 Wiestner A, Schlemper R J, van der Maas A P C, et al. An activating splice donor mutation in the thrombopoietin gene causes hereditary thrombocythaemia. Nat Genet, 1998, 18: 49–52

24 Sonenberg N, Hinnebusch A G. Regulation of translation initiation in eukaryotes: mechanisms and biological targets. Cell, 2009, 136: 731–745

25 Pisareva V P, Pisarev A V. DHX29 reduces leaky scanning through an upstream AUG codon regardless of its nucleotide context. Nucleic Acids Res, 2016, 44: 4252–4265

26 Andreev D E, O'Connor P B F, Fahey C, et al. Translation of 5′ leaders is pervasive in genes resistant to eIF2 repression. eLife, 2015, 4: e03971

27 Medenbach J, Seiler M, Hentze M W. Translational control via protein-regulated upstream open reading frames. Cell, 2011, 145: 902–913

28 Johnstone T G, Bazzini A A, Giraldez A J. Upstream ORFs are prevalent translational repressors in vertebrates. EMBO J, 2016, 35: 706–723

29 Kervestin S, Jacobson A. NMD: a multifaceted response to premature translational termination. Nat Rev Mol Cell Biol, 2012, 13: 700–712

30 He F, Li X, Spatrick P, et al. Genome-wide analysis of mRNAs regulated by the nonsense-mediated and 5′ to 3′ mRNA decay pathways in yeast. Mol Cell, 2003, 12: 1439–1452

31 Hurt J A, Robertson A D, Burge C B. Global analyses of UPF1 binding and function reveal expanded scope of nonsense-mediated mRNA decay. Genome Res, 2013, 23: 1636–1650

32 González C I, Bhattacharya A, Wang W, et al. Nonsense-mediated mRNA decay in *Saccharomyces cerevisiae*. Gene, 2001, 274: 15–25

33 He F, Brown A H, Jacobson A. Upf1p, Nmd2p, and Upf3p are interacting components of the yeast nonsense-mediated mRNA decay pathway.. Mol Cell Biol, 1997, 17: 1580–1594

34 Ruiz-Echevarría M J, Peltz S W. The RNA binding protein Pub1 modulates the stability of transcripts containing upstream open reading frames. Cell, 2000, 101: 741–751

35 Kim J M, Jang S A, Yu B J, et al. High-level expression of an antimicrobial peptide histonin as a natural form by multimerization and furin-mediated cleavage. Appl Microbiol Biotechnol, 2008, 78: 123–130

36 Finley D, Ulrich H D, Sommer T, et al. The ubiquitin-proteasome system of *Saccharomyces cerevisiae*. Genetics, 2012, 192: 319–360

37 Bodzon-Kulakowska A, Bierczynska-Krzysik A, Dylag T, et al. Methods for samples preparation in proteomic research. J Chromatogr B, 2007,

849: 1–31
38 Rabilloud T. Solubilization of proteins for electrophoretic analyses. Electrophoresis, 1996, 17: 813–829
39 Görg A, Weiss W, Dunn M J. Current two-dimensional electrophoresis technology for proteomics. Proteomics, 2004, 4: 3665–3685
40 Broeckx V, Boonen K, Pringels L, et al. Comparison of multiple protein extraction buffers for GeLC-MS/MS proteomic analysis of liver and colon formalin-fixed, paraffin-embedded tissues. Mol Biosyst, 2016, 12: 553–565
41 Ma J, Diedrich J K, Jungreis I, et al. Improved identification and analysis of small open reading frame encoded polypeptides. Anal Chem, 2016, 88: 3967–3975
42 Wilkinson D, Ramsdale M. Proteases and caspase-like activity in the yeast *Saccharomyces cerevisiae*. Biochem Soc Trans, 2011, 39: 1502–1508
43 Che F Y, Zhang X, Berezniuk I, et al. Optimization of neuropeptide extraction from the mouse hypothalamus. J Proteome Res, 2007, 6: 4667–4676
44 Hu L, Li X, Jiang X, et al. Comprehensive peptidome analysis of mouse livers by size exclusion chromatography prefractionation and nanoLC-MS/MS identification. J Proteome Res, 2007, 6: 801–808
45 Hölttä M, Zetterberg H, Mirgorodskaya E, et al. Peptidome analysis of cerebrospinal fluid by LC-MALDI MS. PLoS ONE, 2012, 7: e42555
46 Dallas D C, Guerrero A, Khaldi N, et al. Extensive *in vivo* human milk peptidomics reveals specific proteolysis yielding protective antimicrobial peptides. J Proteome Res, 2013, 12: 2295–2304
47 Manes N P, Gustin J K, Rue J, et al. Targeted protein degradation by Salmonella under phagosome-mimicking culture conditions investigated using comparative peptidomics. Mol Cell Proteomics, 2007, 6: 717–727
48 Khmelnitsky Y L, Belova A B, Levashov A V, et al. Relationship between surface hydrophilicity of a protein and its stability against denaturation by organic solvents. FEBS Lett, 1991, 284: 267–269
49 Polson C, Sarkar P, Incledon B, et al. Optimization of protein precipitation based upon effectiveness of protein removal and ionization effect in liquid chromatography-tandem mass spectrometry. J Chromatogr B, 2003, 785: 263–275
50 Zhang F, Chen J Y. A method for identifying discriminative isoform-specific peptides for clinical proteomics application. BMC Genomics, 2016, 17: 522
51 Schägger H. Tricine-SDS-PAGE. Nat Protoc, 2006, 1: 16–22
52 Hao F, Li J, Zhai R, et al. A novel microscale preparative gel electrophoresis system. Analyst, 2016, 141: 4953–4960
53 Xu Y, Cao Q, Svec F, et al. Porous polymer monolithic column with surface-bound gold nanoparticles for the capture and separation of cysteine-containing peptides. Anal Chem, 2010, 82: 3352–3358
54 Foettinger A, Leitner A, Lindner W. Selective enrichment of tryptophan-containing peptides from protein digests employing a reversible derivatization with malondialdehyde and solid-phase capture on hydrazide beads. J Proteome Res, 2007, 6: 3827–3834
55 Grunert T, Pock K, Buchacher A, et al. Selective solid-phase isolation of methionine-containing peptides and subsequent matrix-assisted laser desorption/ionisation mass spectrometric detection of methionine- and of methionine-sulfoxide-containing peptides. Rapid Commun Mass Spectrom, 2003, 17: 1815–1824
56 Thingholm T E, Jørgensen T J D, Jensen O N, et al. Highly selective enrichment of phosphorylated peptides using titanium dioxide. Nat Protoc, 2006, 1: 1929–1935
57 Zhang L, Zhao Q, Liang Z, et al. Synthesis of adenosine functionalized metal immobilized magnetic nanoparticles for highly selective and sensitive enrichment of phosphopeptides. Chem Commun, 2012, 48: 6274–6276
58 Wada Y, Tajiri M, Yoshida S. Hydrophilic affinity isolation and MALDI multiple-stage tandem mass spectrometry of glycopeptides for glycoproteomics. Anal Chem, 2004, 76: 6560–6565
59 Neue K, Mormann M, Peter-Katalinic J, et al. Elucidation of glycoprotein structures by unspecific proteolysis and direct nanoESI mass spectrometric analysis of ZIC-HILIC-enriched glycopeptides. J Proteome Res, 2011, 10: 2248–2260
60 Choudhary G, Wu S L, Shieh P, et al. Multiple enzymatic digestion for enhanced sequence coverage of proteins in complex proteomic mixtures using capillary LC with ion trap MS/MS. J Proteome Res, 2003, 2: 59–67
61 Hartmann E M, Armengaud J. N-terminomics and proteogenomics, getting off to a good start. Proteomics, 2014, 14: 2637–2646
62 Fischer F, Poetsch A. Protein cleavage strategies for an improved analysis of the membrane proteome. Proteome Sci, 2006, 4: 2
63 Min L, Choe L H, Lee K H. Improved protease digestion conditions for membrane protein detection. Electrophoresis, 2015, 36: 1690–1698
64 Lin Y, Zhou J, Bi D, et al. Sodium-deoxycholate-assisted tryptic digestion and identification of proteolytically resistant proteins. Anal Biochem,

2008, 377: 259–266
65 Wiśniewski J R, Zougman A, Nagaraj N, et al. Universal sample preparation method for proteome analysis. Nat Meth, 2009, 6: 359–362
66 Whitelegge J. Intact protein mass spectrometry and top-down proteomics. Expert Rev Proteomics, 2013, 10: 127–129
67 Sleator R D. An overview of the current status of eukaryote gene prediction strategies. Gene, 2010, 461: 1–4
68 Cheng H, Chan W S, Li Z, et al. Small open reading frames: current prediction techniques and future prospect. Curr Protein Peptide Sci, 2011, 12: 503–507
69 Yates J R, Eng J K, McCormack A L. Mining genomes: correlating tandem mass spectra of modified and unmodified peptides to sequences in nucleotide databases. Anal Chem, 1995, 67: 3202–3210
70 Roos F F, Jacob R, Grossmann J, et al. PepSplice: cache-efficient search algorithms for comprehensive identification of tandem mass spectra. Bioinformatics, 2007, 23: 3016–3023
71 Specht M, Stanke M, Terashima M, et al. Concerted action of the new Genomic Peptide Finder and AUGUSTUS allows for automated proteogenomic annotation of the *Chlamydomonas reinhardtii* genome. Proteomics, 2011, 11: 1814–1823
72 Ning K, Nesvizhskii A I. The utility of mass spectrometry-based proteomic data for validation of novel alternative splice forms reconstructed from RNA-Seq data: a preliminary assessment. BMC Bioinformatics, 2010, 11: S14
73 Van Damme P, Gawron D, Van Criekinge W, et al. N-terminal proteomics and ribosome profiling provide a comprehensive view of the alternative translation initiation landscape in mice and men. Mol Cell Proteomics, 2014, 13: 1245–1261
74 Blakeley P, Overton I M, Hubbard S J. Addressing statistical biases in nucleotide-derived protein databases for proteogenomic search strategies. J Proteome Res, 2012, 11: 5221–5234
75 Fermin D, Allen B B, Blackwell T W, et al. Novel gene and gene model detection using a whole genome open reading frame analysis in proteomics. Genome Biol, 2006, 7: R35
76 Brosch M, Saunders G I, Frankish A, et al. Shotgun proteomics aids discovery of novel protein-coding genes, alternative splicing, and “resurrected” pseudogenes in the mouse genome. Genome Res, 2011, 21: 756–767
77 Branca R M M, Orre L M, Johansson H J, et al. HiRIEF LC-MS enables deep proteome coverage and unbiased proteogenomics. Nat Methods, 2014, 11: 59–62
78 Fu Y, Qian X. Transferred subgroup false discovery rate for rare post-translational modifications detected by mass spectrometry. Mol Cell Proteomics, 2014, 13: 1359–1368
79 Zhang K, Fu Y, Zeng W F, et al. A note on the false discovery rate of novel peptides in proteogenomics. Bioinformatics, 2015, 31: 3249–3253

Advances in small protein identification

HE CuiTong[1,2], ZHANG Yao[2,3] & XU Ping[1,2]

1 Graduate School, Anhui Medical University, Hefei 230032, China;
2 Beijing Proteome Research Center, National Center for Protein Sciences (Beijing), State Key Laboratory of Proteomics, Beijing Institute of Lifeomics, Beijing 102206, China;
3 School of Life Sciences, Sun Yat-sen University, Guangzhou 510275, China

Small proteins (SPs) are generally encoded by short open reading frame (sORF), containing 100 amino acids or less. Recent research has showed that it is required for many crucial biological events, including signal transduction, metabolism, growth and so on. But for the research of SPs, there are many challenges because of the limitation of genome annotation and biochemical detection technologies. The highly effective SPs appraisal or the identification technology is one of essential foundations for its functional research and genomic perfection. This review highlights the difficulties, causes and solutions in SPs identification.

proteomics, small protein, identification

doi: 10.1360/N052017-00245