蛋白质组技术的研究进展

大规模基因组测序计划的实施已改变生命科学的重心，在相当短的时期内，一些原核生物和某些低等真核生物的基因组序列已被测定. 1995 年，流感嗜血杆菌基因组序列首次被破译，在此后不到两年的时间，近 50 个细菌的基因组序列已被完成. 然而，这仅仅是理解有机物功能的一个起点. 在基因组时代，许多 DNA 序列信息仅提供相关基因组的结构和功能. 然而，对基因产物（mRNA 和蛋白质）的理解是理解细胞生物学的一个不可缺少的部分. DNA 序列信息不能预测：1)基因表达产物是否或何时被翻译；2)基因产物的相应含量；3)翻译后修饰的程度；4)基因剔除或过表达的影响；5)遗留的小基因或<300 bp 的 ORFs 的出现；6)多基因现象的表型. 此外，mRNA 水平的测量并不能完全揭示细胞调节；且蛋白质的样品较 mRNA 稳定；蛋白质和 mRNA 之间的相关系数仅为 0.4～0.5，还存在转录后加工、翻译调节以及翻译后加工等. 故而，“基因组时代”的迅猛发展同时激起了人们对“后基因组时代”中蛋白质组研究的需求.

1 蛋白质组的含义

蛋白质组(Proteome)的概念最先由 Marc Wilkins[1]提出，指由一个基因组(genOME)，或一个细胞、组织表达的所有蛋白质(PROTein). 蛋白质组的概念与基因组的概念有许多差别，它随着组织、甚至环境状态的不同而改变. 在转录时，一个基因可以多种 mRNA 形式剪接，并且，同一蛋白可能以许多形式进行翻译后的修饰[2]. 故一个蛋白质组不是一个基因组的直接产物，蛋白质组中蛋白质的数目有时可以超过基因组的数目[3].

蛋白质组学(Proteomics)处于早期“发育”状态，这个领域的专家否认它是单纯的方法学[4]，就像基因组学一样，不是一个封闭的、概念化的稳定的知识体系，而是一个领域. 蛋白质组学集中于动态描述基因调节，对基因表达的蛋白质水平进行定量的测定，鉴定疾病、药物对生命过程的影响，以及解释基因表达调控的机制[5]. 作为一门科学，蛋白质组研究并非从零开始，它是已有 20 年历史的蛋白质(多肽)谱和基因产物图谱技术的一种延伸. 多肽图谱依靠双向电泳(Two-dimensional gel electrophoresis, 2-DE)和进一步的图象分析；而基因产物图谱依靠多种分离后的分析，如质谱技术、氨基酸组分分析等.

2 蛋白质组研究的核心 用于分离的双向电泳(2-DE)

蛋白质组研究的发展以双向电泳技术作为核心. 双向电泳由 O'Farrell's[6]于 1975 年首次建立并成功地分离约 1 000 个 *E.coli* 蛋白，并表明蛋白质谱不是稳定的，而是随环境而变化. 双向电泳原理简明，第一向进行等电聚焦，蛋白质沿 pH 梯度分离，至各自的等电点；随后，再沿垂直的方向进行分子量的分离. 目前，随着技术的飞速发展，已能分离出 10 000 个斑点(spot)[7]. 当双向电泳斑点的全面分析成为现实的时候，蛋白质组的分析变得可行.

样品制备(sample prepareation)和溶解同样事关 2-DE 的成效，目标是尽可能扩大其溶解度和解聚，以提高分辨率. 用化学法和机械裂解法破碎以尽可能溶解和解聚蛋白，两者联合有协同作用. 对 IEF(isoelectric focusing)样品的预处理涉及溶解、变性和还原来完全破坏蛋白间的相互作用，并除去如核酸等非蛋白物质. 理想的状态是人们应一步完成蛋白的完全处理[2]. 近来， 在"变性剂鸡尾酒"中，含 14～16 个碳的磺基甘氨酸三甲内盐($ASB_{14\sim16}$)的裂解液效果最好[8]. 而离液剂 2 mol/L 硫脲和表面活性剂 4%CHAPS 的混合液促使疏水蛋白从 IPG (immobilized pH gradients)胶上的转换[9]. 三丁基膦(Tributyl phosphine，TBP)取代β-巯基乙醇或 DTT 完全溶解链间或链内的二硫键，增强了蛋白的溶解度，并导致转至第二向的增加[10]. 两者通过不同的方法来增加蛋白的溶解度，作为互补试剂会更有效. 在保持样品的完整性的前提下，可利用超离[2]和核酸内切酶[11]去除核酸(DNA). 除此之外，机械力被用来对蛋白分子解聚，如超声破碎[12]等. 另外，添加 PMSF 等蛋白酶抑制剂[13]，可保持蛋白完整性. 由于商品化的 IPG 胶条是干燥脱水的，可在其水化的过程中加样，覆盖整个 IPG 胶，避免在样品杯中的沉淀所致的样品丢失[14]. 此外，低丰度蛋白(low abundance protein)在细胞内可能具有重要的调节功能，代表蛋白质组研究的"冰山之尖"，故分离低丰度蛋白是一种挑战[15]. 亚细胞分级和蛋白质预分级、提高加样量(已达到 1～15 mg 级的标准)[16, 17]、应用敏感性检测，可以提高其敏感性. 如一种多肽免疫 2-DE 印迹(MI-2DE)是利用几种单克隆抗体技术来分析和检测[18]. 提高组蛋白和核糖体蛋白等碱性蛋白(basic proteins)的分离是另一难点. 由于碱性 pH 范围内凝胶基质的不稳定及逆向电渗流(EOF)的产生，对 PI(等电点)超过 10 的碱性蛋白[11]，通过产生 0～10% 的山梨醇梯度和 16%的异丙醇可减少之. 亦可用双甲基丙烯酰胺来增加基质的稳定性[19].

2-DE 面临的挑战是高分辨率和重复性. 高分辨率确保蛋白最大程度的分离，高重复性允许进行凝胶间配比(match). 对 2-DE 而言，有 3 种方法分离蛋白：1) ISO-DALT(isoelectric focus)以 O'Farrell's 技术为基础. 第一向应用载体两性电解质(carrier ampholyte, CA)，在管胶内建立 pH 梯度. 随着聚焦时间的延长，pH 梯度不稳，易产生阴极漂移. 2) NEPHGE(non-equilibrium pH gradient electrophoresis)[20]用于分离碱性蛋白(pH>7.0). 如果聚焦达到平衡状态，碱性蛋白

会离开凝胶基质而丢失. 因此，在等电区域的迁移须在平衡状态之前完成，但很难控制. 3)IPG-DALT 发展于 80 年代早期. 由于固相 pH 梯度(Immobilized pH gradient, IPG)[21]的出现解决了 pH 梯度不稳的问题. IPG 通过 immobiline 共价偶联于丙烯酰胺产生固定的 pH 梯度，克服了 IEF 的缺点，从而达到高度的重复性. 目前可以精确制作线性、渐进性和 S 型曲线，范围或宽或窄的 pH 梯度. 新的酸性 pH 3～5 或碱性 pH 6～11 的 IPG 凝胶梯度联合商品化的 pH 4～7 的梯度可对蛋白质形成蛋白质组重叠群(proteomic contigs)从而有效分离[13].

分离后的斑点检测(spot detection)亦很重要. 所采用的检测策略和分离后所采用的方法的相互作用是很重要的. 此外，还需考虑反应的线性、饱和阈/动态范围、敏感性、对细胞蛋白群的全体定量分析的适应性、可行性. 目前，没有一种蛋白染色覆盖广泛的浓度和 PI 及分离后分析技术. 银染已成为一种检测 2-DE 的流行方法，可检测少到 2～5ng 的蛋白，因此较考马斯亮蓝 R-250 敏感. 多数糖蛋白不能被考马斯亮蓝染色，一些有机染料不适于 PVDF 膜. 放射性标记不依赖其代谢的活性，并仅适于对合成的蛋白质检测[2]. 另有一种改良的 2-DE(差异凝胶电泳)，即应用两种不同的染料荧光标记两个样品，使在同一凝胶上电泳后的凝胶图象为两个，避免了几种 2-DE 的比较，可在纳克级进行检测[22].

较早期相比，2-DE 有两个主要的进步：首先，极高的重复性使有机体的参考图谱，可通过 Internet 获得，来比较不同组织类型、不同状态的基因表达；其次，高加样量使得 2-DE 成为一项真正的制备型技术.

3 蛋白质组技术的支柱---鉴定技术(Identification)

如果目前分离蛋白质组的最好技术是 2-DE，那么随之而来的挑战是数百数千个蛋白如何被鉴定. 在这里，我们不考虑传统的蛋白鉴定方法，如免疫印迹法、内肽的化学测序、已知或未知蛋白的 comigration 分析，或者在一个有机体中有意义的基因的过表达. 并不是因为这些方法无效，而是因为它们通常耗时、耗力，不适合高流通量的筛选. 目前，所选用的技术包括对于蛋白鉴定的图象分析、微量测序；进一步对肽片段进行鉴定的氨基酸组分分析和与质谱相关的技术.

(1) 图象分析技术(Image analysis). “满天星”式的 2-DE 图谱分析不能依靠本能的直觉，每一个图象上斑点的上调、下调及出现、消失，都可能在生理和病理状态下产生，必须依靠计算机为基础的数据处理，进行定量分析. 在一系列高质量的 2-DE 凝胶产生(低背景染色，高度的重复性)的前提下，图象分析包括斑点检测、背景消减、斑点配比和数据库构建. 首先，采集图象通常所用的系统是电荷耦合 CCD(charge coupled device)照相机；激光密度仪(laser densitometers)和 Phospho 或 Fluoro imagers，对图象进行数字化. 并成为以象素(pixel

s)为基础的空间和网格. 其次，在图象灰度水平上过滤和变形，进行图象加工，以进行斑点检测. 利用 Laplacian，Gaussian，DOG(difference of Gaussians) opreator 使有意义的区域与背景分离，精确限定斑点的强度、面积、周长和方向. 图象分析检测的斑点须与肉眼观测的斑点一致. 在这一原则下，多数系统以控制斑点的重心或最高峰来分析，边缘检测的软件可精确描述斑点外观，并进行边缘检测和邻近分析，以增加精确度. 通过阈值分析、边缘检测、销蚀和扩大斑点检测的基本工具还可恢复共迁移的斑点边界. 以 PC 机为基础的软件 Phoretix-2D 正挑战古老的 Unix 为基础的 2-D 分析软件包. 第三，一旦 2-DE 图象上的斑点被检测，许多图象需要分析比较、增加、消减或均值化. 由于在 2-DE 中出现 100%的重复性是很困难的，由此凝胶间的蛋白质的配比对于图象分析系统是一个挑战. IPG 技术的出现已使斑点配比变得容易. 因此，较大程度的相似性可通过斑点配比向量算法在长度和平行度观测. 用来配比的著名软件系统包括 Quest，Lips，Hermes，Gemini 等，计算机方法如相似性、聚类分析、等级分类和主要因素分析已被采用，而神经网络、子波变换和实用分析在未来可被采用[2]. 配比通常由一个人操作，其手工设定大约 50 个突出的斑点作为“路标”，进行交叉配比. 之后，扩展至整个胶. 例如：精确的 PI 和 MW(分子量)的估计通过参考图上 20 个或更多的已知蛋白所组成的标准曲线来计算未知蛋白的 PI 和 MW[3]. 在凝胶图象分析系统依据已知蛋白质的 pI 值产生 PI 网络，使得凝胶上其它蛋白的 PI 按此分配. 所估计的精确度大大依赖于所建网格的结构及标本的类型. 已知的未被修饰的大蛋白应该作为标志，变性的修饰的蛋白的 PI 估计约在±0.25 个单位. 同理，已知蛋白的理论分子量可以从数据库中计算，利用产生的表观分子量的网格来估计蛋白的分子量. 未被修饰的小蛋白的错误率大约 30%，而翻译后蛋白的出入更大. 故需联合其他的技术完成鉴定[18].

(2) 微量测序(microsequencing). 蛋白质的微量测序已成为蛋白质分析和鉴定的基石，可以提供足够的信息. 尽管氨基酸组分分析和肽质指纹谱(PMF)可鉴定由 2-DE 分离的蛋白，但最普通的 N-末端 Edman 降解仍然是进行鉴定的主要技术. 目前已实现蛋白质微量测序的自动化. 首先使经凝胶分离的蛋白质直接印迹在 PVDF 膜或玻璃纤维膜上，染色、切割，然后直接置于测序仪中，可用于 subpicomole 水平的蛋白质的鉴定[2]. 但有几点需注意：Edman 降解很缓慢，序列以每 40 min 1 个氨基酸的速率产生；与质谱相比，Edman 降解消耗大；试剂昂贵，每个氨基酸花费 3～4$. 这都说明泛化的 Edman 降解蛋白质不适合分析成百上千的蛋白质. 然而，如果在一个凝胶上仅有几个有意义的蛋白质，或者如果其他技术无法测定而克隆其基因是必需的，则需要进行泛化的 Edman 降解测序.

近来，应用自动化的 Edman 降解可产生短的 N-末端序列标签，这是将质谱的序列标签概念用于 Edman 降解，业已成为一种强有力的蛋白质鉴定. 当对 Edman 的硬件进行简单改进，以迅速产生 N-末端序列标签达 10～20 个/d，序列检签将适于在较小的蛋白质组中进行鉴定.若联合其他的蛋白质属性，如氨基酸组分分析、肽质质量、表现蛋白质分子量、等电点，可以更加可信地鉴定蛋白质. 选择 BLAST 程序，可与数据库相配比[18]. 目前，采用一种 TagIdent 的检索程序，还可以进行种间比较鉴定，又提高了其在蛋白质组研究中的作用[23].

(3) 与质谱(mass spectrometry)相关的技术. 质谱已成为连接蛋白质与基因的重要技术，开启了大规模自动化的蛋白质鉴定之门. 用来分析蛋白质或多肽的质谱有两个主要的部分，1)样品入机的离子源，2)测量被介入离子的分子量的装置. 首先是基质辅助激光解吸附电离飞行时间质谱(MALDI-TOF)为一脉冲式的离子化技术. 它从固相标本中产生离子，并在飞行管中测其分子量. 其次是电喷雾质谱(ESI-MS)，是一连续离子化的方法，从液相中产生离子,联合四极质谱或在飞行时间检测器中测其分子量[24]. 近年来，质谱的装置和技术有了长足的进展. 在 MALDI-TOF 中，最重要的进步是离子反射器(ion reflectron)和延迟提取(delayed ion extraction)，可达相当精确的分子量. 在 ESI-MS 中，纳米级电雾源(nano-electrospray source)的出现使得微升级的样品在 30～40 min 内分析成为可能. 将反相液相色谱和串联质谱(tandem MS)联用，可在数十个 picomole 的水平检测；若利用毛细管色谱与串联质谱联用，则可在低 picomole 到高 femtomole 水平检测；当利用毛细管电泳与串联质谱连用时，可在小于 femtomole 的水平检测[25]. 甚至可在 attomole 水平进行[26]. 目前多为酶解、液相色谱分离、串联质谱及计算机算法的联合应用鉴定蛋白质. 下面以肽质指纹术和肽片段的测序来说明怎样通过质谱来鉴定蛋白质[2].

1)肽质指纹术(peptide mass fingerprint, PMF)是由 Henzel 等人[27]于 1993 年提出. 用酶(最常用的是胰酶)对由 2-DE 分离的蛋白在胶上或在膜上于精氨酸或赖氨酸的 C-末端处进行断裂，断裂所产生的精确的分子量通过质谱来测量(MALDI-TOF-MS，或为 ESI-MS)，这一技术能够完成的肽质量可精确到 0.1 个分子量单位. 所有的肽质量最后与数据库中理论肽质量相配比(理论肽是由实验所用的酶来“断裂”蛋白所产生的). 配比的结果是按照数据库中肽片段与未知蛋白共有的肽片段数目作一排行榜，“冠军”肽片段可能代表一个未知蛋白.若冠亚军之间的肽片段存在较大差异，且这个蛋白可与实验所示的肽片段覆盖良好，则说明正确鉴定的可能性较大[18].

2)肽片段(peptide fragment)的部分测序. 肽质指纹术对其自身而言，不能揭示所衍生的肽片段或蛋白质. 为进一步鉴定蛋白质，出现了一系列的质谱方法用

来描述肽片段. 用酶或化学方法从 N-或 C-末端按顺序除去氨基酸，形成梯形肽片段(ladder peptide)[28]. 首先以一种可控制的化学模式从 N-末端降解，可产生大小不同的一系列的梯形肽片段，所得一定数目的肽质量由 MALDI-TOF-MS 测量. 另一种方法涉及羧基肽酶的应用，从 C-末端除去不同数目的氨基酸形成肽片段. 化学法和酶法可产生相对较长的序列，其分子量精确至以区别赖氨酸(128.09)和谷氨酰胺(128.06)[18]. 或者，在质谱仪内应用源后衰变(post-source decay, PSD)和碰撞诱导解离(collision-induced dissociation, CID)，目的是产生包含有仅异于一个氨基酸残基质量的一系列肽峰的质谱. 因此，允许推断肽片段序列[29]. 肽片段 PSD 的分析在 MALDI 反应器上能产生部分序列信息. 首先进行肽质指纹鉴定. 之后，一个有意义的肽片段在质谱仪被选作“母离子”，在飞行至离子反应器的过程中降解为“子离子”. 在反应器中，用逐渐降低的电压可测量至检测器的不同大小的片段[30]. 但经常产生不完全的片段. 现在用肽片段来测序的方法始于 70 年代末的 CID，可以一个三联四极质谱 ESI-MS 或 MALDI-TOF-MS 联合碰撞器内来完成. 在 ESI-MS 中，由电雾源产生的肽离子在质谱仪的第一个四极质谱中测量，有意义的肽片段被送至第二个四极质谱中，惰性气体轰击使其成为碎片，所得产物在第三个四极质谱中测量[31]. 与 MALDI-PSD 相比，CID 稳定、强健、普遍，肽离子片段基本沿着酰胺键的主架被轰击产生梯形序列. 连续的片段间差异决定此序列在那一点的氨基酸的质量. 由此，序列可被推测. 由 CID 图谱还可获得的几个序列的残基，叫做“肽序列标签”. 这样，联合肽片段母离子的分子量和肽片段距 N-、C 端的距离将足以鉴定一个蛋白质[32].

(4) 氨基酸组分分析. 1977 年首次作为鉴定蛋白质的一种工具，是一种独特的“脚印”技术. 利用蛋白质异质性的氨基酸组分特征，成为一种独立于序列的属性，不同于肽质量或序列标签. Latter 首次表明氨基酸组分的数据能用于从 2-DE 凝胶上鉴定蛋白质[33]. 通过放射标记的氨基酸来测定蛋白质的组分，或者将蛋白质印迹到 PVDF 膜上，在 155℃进行酸性水解 1 h，通过这一简单步骤的氨基酸的提取，每一样品的氨基酸在 40min 内自动衍生并由色谱分离，常规分析为 100 个蛋白质/周[3]. 依据代表两组分间数目差异的分数，对数据库中的蛋白质进行排榜，“冠军”蛋白质具有与未知蛋白质最相近的组分，考虑冠亚军蛋白质分数之间的差异，仅处于冠军的蛋白质的可信度大. Internet 上存在多个程序可用于氨基酸组分分析，如 AACompIdent，ASA，FINDER，AAC-PI，PROP-SEARCH 等，其中，在 PROP-SEARCH 中，组分、序列和氨基酸的位置被用来检索同源蛋白质[2]. 但仍存在一些缺点，如由于不足的酸性水解或者部分降解会产生氨基酸的变异. 故应联合其他的蛋白质属性进行鉴定.

4 蛋白质组研究的百科全书 数据库(database)

蛋白质组数据库(proteome database)被认为是蛋白质组知识的储存库，包含所有鉴定的蛋白质信息，如蛋白质的顺序、核苷酸顺序、2-D PAGE、3-D 结构、翻译后的修饰、基因组及代谢数据库等. 例如，SWISS-2DPAGE 数据库包括人类，细菌，细胞等物种的信息[34]. 其中，*E.coli* SWISS-2DPAGE 数据库是 EXPASY 分子生物学服务器的一部分，通过 www 的 URL 网址 http://www.expasy.ch/ch2d/ch2d-top.html[35] 可以查询.

当前的计算机和网络技术，让我们将所有的数据库连在一起，并允许我们从一个数据库中的一条信息遨游到其他的数据库；将一个研究对象的数据与其他各种蛋白质组中的相关数据或图谱相连. 分析型软件工具被称为蛋白质组分析机器人、数据分析软件包. 在既定的状态下，定量研究蛋白质的表达水平，或者计算机辅助数据库系统建立可将实验推进一步.因此，蛋白质组分析技术联合蛋白质数据库，计算机网络和其他软件包合在一起称为蛋白质组的机控百科全书(Cyber-encyclopaedia of the proteome)[18].

蛋白质组和基因组共同分析可以产生大量的数据. 当评估每一个数据库的价值时，难免要考虑两个条件：1)数据库是否在任一时刻保持最新；2)何时能够相互连接，且以整体状态评估. 目前的发展趋势：1)信息量呈指数增长；2)蛋白质组计划的实施会产生新的数据库；3)致力于模拟细胞内蛋白质的相互作用的新型数据库；4)建立高级、智慧型的咨询工具是必需的［[18].

5 蛋白质组技术的规模 高流通量筛选（HTS）

HTS(High throughput screening)至今在蛋白质组研究中已成为现实. 在最近的一年内，由于制药工业对此的需求，样品输入自动化得以进展. 目前，正在设计的机器人可自动处理 2-DE 后电转至 PVDF 膜. 原形机器人加工、传输蛋白质至质谱或以液相色谱为基础的分析仪，如进行斑点切割，操纵、控制多种 PMF、氨基酸组分分析所需的化学反应，使每天最小的流通量达 1000 个蛋白. 此外，必须选择适用的软件包，如应用第二代 COMBINED 来处理输出的数据，自动咨询本地或网上的数据库而进行系列的评估. 大量的数据分析表明 HTS 是刻不容缓的. 目前，对质谱已设想一个三级方案来处理大规模的蛋白质组：1）MALDI-TOF-MS 以每天大于 1000 个蛋白的速率分析；2）通过 ESI-MS/MS 或 SEQUEST，以每天每台机器分析几打蛋白质的速率进行序列标签；3）对由串联质谱所得的新蛋白或有意义蛋白进行全长肽段的测序，从而提供足够的信息通过核酸探针或简并 PCR 引物获得有意义的基因[2,36].

综上所述, 高分辨率、高敏感性和高流通性的分离和分离后鉴定技术, 结合准确、全面的数据库技术, 使蛋白质组技术用于生物研究卓有成效. 但仅鉴定蛋白

质是不够的，蛋白质组世界的挑战是完善蛋白质质和量的分析,设想细胞活性、功能的全体性概念. 在此基础上,蛋白质组分析将会促进未来生命科学的整体发展.

地址：杭州市西湖科技园西园八路 11 号
邮编：310030
售后服务专线：400-672-1817
销售电话：0571-86056609 86059660
86054117 86055117
传真：0571-86059660 86823529
网址：www.top17.net