

生物科学研究方法丛书

蛋白质组学方法

中国生物技术的发展中心
南开大学

编著



科学出版社

生物科学研究方法丛书

蛋白质组学方法

中国科学院上海生物化学研究所

潘蔚 主编

科学出版社

生物科学研究方法丛书

蛋白质组学方法

中国生物技术发展中心 编著
南开大学

主 编 饶子和

副主编 沈月全

主要参编人员 (按姓氏汉语拼音排序)

陈凌懿	龚清秋	韩际宏	洪章勇
胡俊杰	旷 苗	刘新奇	龙加福
娄志勇	门淑珍	邱宏伟	饶子和
沈月全	王 晶	王 莹	吴世安
杨志谋	赵饮虹	周卫红	

科 学 出 版 社

北 京

· 版权所有 侵权必究 ·

举报电话:010-64030229;010-64034315;13501151303(打假办)

内 容 简 介

本书总结和归纳了现代生物学领域内的常用蛋白质研究技术。详细阐述了这些技术的原理、方法和实际应用。使读者能够快速了解这些技术的应用范围,从而选择合适的技术进行科学实验。本书还对这些技术的历史进行了回顾,并在此基础上,对将来技术的发展提出一些见解和看法。

本书对研究人员进一步发展这些技术具有一定的指导意义。

图书在版编目(CIP)数据

蛋白质组学方法 / 中国生物技术的发展中心,南开大学编著;饶子和主编. —北京:科学出版社,2012.9

(生物科学研究方法丛书)

ISBN 978-7-03-035584-3

I. 蛋… II. ①中… ②南… ③饶… III. 蛋白质-基因组-研究方法
IV. Q51-3

中国版本图书馆 CIP 数据核字(2012)第 221819 号

责任编辑:胡治国 邹梦娜 / 责任校对:林青梅

责任印制:肖 兴 / 封面设计:范璧合

版权所有,违者必究。未经本社许可,数字图书馆不得使用

科学出版社 出版

北京东黄城根北街16号

邮政编码:100717

<http://www.sciencep.com>

新科印刷有限公司 印刷

科学出版社发行 各地新华书店经销

*

2012年9月第 一 版 开本:787×1092 1/16

2012年9月第一次印刷 印张:13 1/2

字数:319 000

定价:59.00 元

(如有印装质量问题,我社负责调换)

前 言

随着人类基因组计划的实施和推进,生命科学研究已进入了后基因组时代。在这个时代,生命科学的主要研究对象已经从基因组转变为蛋白质组。尽管现在已有多个物种的基因组被测序,但在这些基因组中通常有一半以上基因的功能是未知的。而蛋白质是生理功能的执行者,是生命现象的直接体现者,对蛋白质结构和功能的研究将更有助于我们直接阐明生命在生理或病理条件下的变化机制。蛋白质本身的存在形式和活动规律,如翻译后修饰、蛋白质间相互作用以及蛋白质构象等问题,仍需要我们利用蛋白质组研究技术直接对蛋白质进行研究来解决。虽然蛋白质的可变性和多样性等特殊性质导致了蛋白质研究技术远远比核酸技术要复杂和困难得多,但正是这些特性参与并影响着整个生命过程。因此在20世纪90年代中期,国际上产生了一门新兴学科——蛋白质组学(proteomics),它是以细胞内全部蛋白质的存在及其活动方式为研究对象。可以说蛋白质组研究的开展不仅是生命科学研究进入后基因组时代的里程碑,也是后基因组时代生命科学研究的核心内容之一。

可以说,研究方法既可以推动蛋白质组学的发展也可以限制其发展。蛋白质组学研究成功与否,速度快慢,很大程度上取决于对其研究方法水平的高低。蛋白质研究方法远比基因复杂和困难:不仅氨基酸残基种类远多于核苷酸残基(20/4),而且蛋白质有着复杂的翻译后修饰,如磷酸化和糖基化等,给分离和研究蛋白质带来很多困难。另外,蛋白质的体外表达和纯化也并非易事,从而难以获得大量的目标蛋白。蛋白质组的研究实质上是在细胞水平上对蛋白质进行大规模的平行分离和分析,往往要同时处理成千上万种蛋白质。因此,发展高通量、高灵敏度、高准确性的研究方法和技术平台是现在乃至相当一段时间内蛋白质组学研究中的重点和难点。

编 者

2012年8月20日

目 录

前言

第一章 概述	(1)
第一节 分学科创新方法范围	(1)
第二节 分学科创新方法的发展历程	(5)
第三节 分学科创新方法的发展趋势	(47)
第二章 主要创新方法(基本原理和改进方向)	(54)
第一节 各种异源蛋白表达技术	(54)
第二节 蛋白质组织抽提技术	(57)
第三节 蛋白质亲和层析与各种色谱纯化技术	(61)
第四节 蛋白质变复性技术	(65)
第五节 二维凝胶电泳与分析技术	(72)
第六节 MALDI-TOF-MS 技术	(78)
第七节 LC-ESI-MS/MS 技术	(80)
第八节 PMF 鉴定技术	(84)
第九节 同位素标记定量分析质谱技术	(85)
第十节 多维液相色谱技术	(88)
第十一节 酵母双杂交技术	(91)
第十二节 免疫共沉淀技术	(94)
第十三节 细胞共定位技术	(95)
第十四节 荧光共振能量转移技术	(96)
第十五节 亲合捕获技术	(98)
第十六节 核酸微阵列及其分析技术	(100)
第十七节 蛋白质芯片及分析技术	(104)
第十八节 蛋白质 N 端和 C 端测序技术	(107)
第十九节 蛋白质 X 射线衍射技术	(113)
第二十节 蛋白质核磁共振及其样品制备和结构测定技术	(125)
第二十一节 单分子技术	(134)
第二十二节 电镜负染观测技术	(139)
第二十三节 蛋白质电子衍射及其二维晶体生长技术	(141)
第二十四节 冷冻电镜观测及其样品制备和三维重构技术	(143)
第二十五节 电子层析技术	(146)
第二十六节 荧光光谱技术	(149)
第二十七节 圆二色谱技术	(151)
第二十八节 拉曼光谱技术	(154)

第二十九节 等温滴定量热技术	(157)
第三十节 扫描隧道显微镜技术	(159)
第三十一节 表面等离子共振技术	(160)
第三十二节 小角散射结构分析技术	(162)
第三十三节 动态光散射技术	(164)
第三十四节 噬菌体表面展示抗体技术	(165)
第三章 分学科创新方法改进的发展策略和途径	(169)
第一节 我国该学科方法创新的研究基础	(169)
第二节 我国该学科创新方法的研究需求	(182)
第三节 我国该学科方法创新的目标、方向和重点	(194)

第一节 分学科创新方法范围

一、后基因组时代生物科学发展的新形势

生物学是一门研究生命现象及其活动规律的科学。近几年,随着生物学的迅猛发展以及研究在分子层面上的切入,人们从微观角度对生物体有了更加系统和深入的认识。19世纪后期到20世纪50年代初,是现代分子生物学开始产生的阶段。在这一阶段产生了两点对生命本质的重大突破。第一点明确了蛋白质是生命的主要基础物质。19世纪末 Buchner 兄弟证明酵母无细胞提取液能使糖发酵产生乙醇,并且第一次提出酶(enzyme)的名称,酶实际是生物意义的化学催化剂。20世纪20~40年代提纯和结晶了一些酶并且证明了酶的本质是蛋白质。随后其他科学家也陆续证明生命的许多基本现象是由酶和蛋白质来完成的。在此期间对蛋白质结构也进行了探索。1902年蛋白质结构被证明是多肽组装的。20世纪40年代末,Sanger创立二硝基氟苯以及 Edman 发展异硫氰酸苯酯法分析肽链 N 端氨基酸。1953年 Sanger 和 Thompson 完成了第一个多肽分子——胰岛素 A 链和 B 链的氨基酸全序列分析。随着 X 射线蛋白质晶体衍射技术的发展,1950年 Pauling 和 Corey 提出了 α -角蛋白的 α 螺旋结构模型^[1]。第二点确定了生物遗传的物质基础是 DNA。20世纪20~30年代已确认自然界有 DNA 和 RNA 两类核酸,并阐明了核苷酸的组成。由于当时对核苷酸和碱基的定量分析不够精确,得出 DNA 中 A、G、C、T 含量是大致相等的结果,因而曾长期错误地认为 DNA 结构是“四核苷酸”单位的重复,不具有多样性,不能携带更多的信息。后来,更多的实验事实使人们对核酸的功能和结构两方面的认识都有了长足的进步。1944年 O. T. Avery 等证明了肺炎球菌转化因子是 DNA。1952年 A. D. Hershey 和 M. Chase 用同位素³⁵S 和³²P 分别标记 T2 噬菌体的蛋白质和核酸,并对大肠埃希菌进行感染实验,进一步证明了 DNA 是遗传物质。在对 DNA 结构的研究上,1949~1952年 S. Furber 等的 X 射线衍射分析阐明了核苷酸并非平面的空间构象,提出了 DNA 是螺旋结构。1948~1953年 Chargaff 等用新的层析和电泳技术分析组成 DNA 的碱基和核苷酸量,积累了大量的数据,提出了 DNA 碱基组成 A=T、G=C 的 Chargaff 规则,为碱基配对的 DNA 结构认识打下了基础^[2]。

随后现代分子生物学进入了建立和发展阶段,这一阶段是从20世纪50年代初到70年代初。1953年以 Watson 和 Crick 提出的 DNA 双螺旋结构模型作为现代分子生物学诞生的里程碑,开创了分子遗传学基本理论建立和发展的黄金时代。DNA 双螺旋发现的最深刻意义在于:确立了核酸作为信息分子的结构基础;提出了碱基配对是核酸复制、遗传信息传递的基本方式;从而最后确定了核酸是遗传的物质基础,为认识核酸与蛋白质的关系及其在生命中的作用打下了最重要的基础。在此期间的主要进展包括:①遗传信息传递中心

法则的建立。在发现 DNA 双螺旋结构同时, Watson 和 Crick 就提出 DNA 复制的可能模型。其后在 1956 年 A. Kornberg 首先发现 DNA 聚合酶; 1958 年 Meselson 及 Stahl 用同位素标记和超速离心分离实验为 DNA 半保留模型提出了证明; 1968 年 Okazaki 提出 DNA 不连续复制模型; 1972 年 DNA 复制开始需要 RNA 作为引物被证实; 20 世纪 70 年代初获得 DNA 拓扑异构酶, 并对真核 DNA 聚合酶特性做了分析研究; 这些都逐渐完善了对 DNA 复制机制的认识。在研究 DNA 复制将遗传信息传给子代的同时, 提出了 RNA 在遗传信息传到蛋白质过程中起着中介作用的假说。1958 年 Weiss 及 Hurwitz 等发现依赖于 DNA 的 RNA 聚合酶; 1961 年 Hall 和 Spiegelman 用 RNA-DNA 杂交证明 mRNA 与 DNA 序列互补; 逐步阐明了 RNA 转录合成的机制。在此同时认识到蛋白质是接受 RNA 的遗传信息而合成的。20 世纪 50 年代初 Zamecnik 等在形态学和分离的亚细胞组分实验中已发现微粒体(microsome)是细胞内蛋白质合成的部位; 1957 年 Hoagland、Zamecnik 及 Stephenson 等分离出 tRNA 并对它们在合成蛋白质中转运氨基酸的功能提出了假设; 1961 年 Brenner 及 Gross 等观察了在蛋白质合成过程中 mRNA 与核糖体的结合; 1965 年 Holley 首次测出了酵母丙氨酸 tRNA 的一级结构; 特别是在 20 世纪 60 年代 Nirenberg、Ochoa 以及 Khorana 等几组科学家共同努力破译了 RNA 上编码合成蛋白质的遗传密码, 随后研究表明这套遗传密码在生物界具有通用性, 从而认识了蛋白质翻译合成的基本过程。上述重要发现共同建立了以中心法则为基础的分子遗传学基本理论体系。1970 年 Temin 和 Baltimore 又同时从鸡肉瘤病毒颗粒中发现以 RNA 为模板合成 DNA 的逆转录酶, 又进一步补充和完善了遗传信息传递的中心法则。②对蛋白质结构与功能的进一步认识。1956~1958 年 Anfinsen 和 White 根据对酶蛋白的变性和复性实验, 提出蛋白质的三维空间结构是由其氨基酸序列来确定的。1958 年 Ingram 证明正常的血红蛋白与镰刀状细胞溶血症病人的血红蛋白之间, 亚基的肽链上仅有一个氨基酸残基的差别, 使人们对蛋白质一级结构影响功能有了深刻的印象。与此同时, 对蛋白质研究的手段也有改进, 1969 年 Weber 开始应用聚丙烯酰胺凝胶电泳测定蛋白质分子量; 20 世纪 60 年代先后分析得到血红蛋白、核糖核酸酶 A 等一批蛋白质的一级结构; 1973 年氨基酸序列自动测定仪问世。中国科学家在 1965 年人工合成了牛胰岛素; 在 1973 年用 $1.8\text{\AA}$ X 射线衍射分析法测定了牛胰岛素的空间结构, 为认识蛋白质的结构做出了重要贡献^[3]。

而现在已经初步认识生命本质并进入开始改造生命的深入发展阶段。20 世纪 70 年代后, 以基因工程技术的出现作为新的里程碑, 标志着人类深入认识生命本质并能改造生命的新时期开始。目前分子生物学已经从研究单个基因发展到研究生物整个基因组的结构与功能。1977 年 Sanger 测定了 ΦX174 -DNA 全部 5375 个核苷酸的序列; 1978 年 Fiers 等测出 SV-40DNA 全部 5224 对碱基序列; 20 世纪 80 年代 λ 噬菌体 DNA 全部 48 502 碱基对的序列全部测出; 一些小的病毒包括乙型肝炎病毒、艾滋病毒等基因组的全序列也陆续被测定; 1996 年年底许多科学家共同努力测出了大肠埃希菌基因组 DNA 的全序列长 4×10^6 碱基对。测定一个生物基因组核酸的全序列无疑对理解这一生物的生命信息及其功能有极大的意义。1990 年人类基因组计划(Human Genome Project)开始实施, 这是生命科学领域有史以来全球性最庞大的研究计划, 基因组的研究催生了以大规模数据产生为基本特征的生物“组学”, 同时也催生了大规模高通量的生物学研究平台, 当从大量的测序数据发展到大量基因功能数据时, 后基因组时代就来到了, 而以功能数据为基础的、系统研究现象的系统生物学也诞生了^[4]。这种研究的特点是: 以假说为出发点、以数据为基础、综合

各项技术、交叉利用各种学科的知识,视角全面、研究深入。虽然交叉综合的巨大潜力是明显的,但是,今天大规模科学研究中的交叉综合与以往个别科学家之间的合作有很大的操作上的差别与本质上的差别。所谓本质的差别,是信息的量的差别已经导致了研究性质的差别;所谓操作上的差别,是指团队性质的研究与个别实验室之间合作的差别。实现交叉综合的困难在各个层次上都是存在的,国际上功能基因组和系统生物学也还处在初生阶段。

随着大量生物体全基因组序列的获得,特别是人类基因组序列精确图的完成,人们发现仅从基因组序列的角度根本无法完整、系统地阐明生物体的功能,甚至无法确认其中基因的数量、种类乃至多数预测基因的客观存在。众多生命现象之谜,不能直接从基因组序列中得到解答。由于翻译调控和翻译后修饰的存在,mRNA亦不能直接反映细胞内蛋白质的水平及其功能状态;对酵母 mRNA 数据的分析的结果表明:二者的相关性低于 0.5。面对基因组计划所产生的成千上万未曾确认的基因,面对基因组、转录组研究对众多生命现象之谜认识上所谓留下的“真空”,蛋白质组学应运而生^[4]。蛋白质组研究是功能基因组这一前沿研究的战略制高点,是孕育未来生命科学与生物技术重大突破的温床,将成为新世纪最大战略资源争夺的主要“战场”。

人体组织、器官的蛋白质组研究,是人类基因组计划实施所促成的生命科学自然的也是必然的引申和发展。基因组测序的完成,使我们获得破译全部蛋白质的密码,为蛋白质组研究提供了依据和路标;但基因组研究如果不发展到蛋白质组的研究,它所获得的大量一维信息对诠释生命奥秘的价值是很有限的,因为只有蛋白质组研究,才能获得基因组表达产物——蛋白质在时间上与空间上是如何演绎生命的信息。

2001 年,国际人类蛋白质组组织(Human Proteome Organization, HUPO)成立,同时提出了人类蛋白质组计划(HLPP),引起全世界的关注。如同当年人类基因组计划一样,人类蛋白质组计划的提出,将成为生命科学新的发展引擎和前沿。由于“蛋白质是生命的存在方式”,是生命现象的执行者,从基因组计划发展到蛋白质计划就是从探索生命的密码发展到探究生命体本身,它更直接地揭示生命现象特别是人体健康与疾病的机制。人类蛋白质组计划的实施,其意义和对人类的影响将要超过人类基因组计划。

二、面向生产实践

蛋白质组研究所包容的技术范围远大于基因组研究,由于被誉为构筑生命大厦“砖块”的蛋白质中蕴藏着开发疾病诊断方法和新型药物靶标的“钥匙”,这一研究领域一经出现就很快给许多行业包括生物技术、计算机技术、分析仪器、生物信息学、生化试剂等,尤其是制药业带来了无限的商机,使得这一源于生命科学研究的主题迅速立体化,进而产业化。主导企业与相关行业纷纷交叉联盟,带来巨额融资招募精英、集成技术框架,短短几年间规模可观的研发机构不断涌现,使蛋白质组这个新兴产业在全世界范围内如火如荼地展开。

基于蛋白质组较之基因组具有更直接的功能含义,加之其与人类疾病诊断和药物研发更直接的联系,大力发展蛋白质组学及其相关技术可以大大加快发现可以用作药物和治疗的高特异性新靶点的速度,为研制包括生物技术药物在内的新的药物以及生物技术产业化打下良好基础。

蛋白质组学领域的生物技术产业化主要是根据药物靶标的结构生物学进行的研究和药物设计。以结构生物学为基础,对正常生理过程中及与疾病相关的重要蛋白质的结构和

功能进行系统的研究与分析,得到蛋白质药靶的三维结构,进行药物与靶标蛋白相互作用的动力学模拟研究。同时充分利用现有药物分子或中草药有效成分,经修饰和优化,设计具有更高活性的先导化合物^[5]。

以蛋白质组学为基础发现药物靶标研究表明,人体内可能存在的药物作用靶标约有3000~15 000个,而统计结果显示,目前发现的药物靶标不到500个,这说明还有大量的药物作用靶标未被发现。大多数药物靶标都是在生命活动中扮演重要角色的蛋白质,如酶、受体、激素等。通过蛋白质组学的方法比较疾病状态和正常生理状态下蛋白质表达的差异,就有可能找到有效的药物作用靶标,其中应用较多的是二维凝胶电泳(2-DE)和质谱分析(MS)技术。在2-DE中,蛋白质样品根据其等电点和相对分子质量的不同而分离,在得到的电泳图谱中,疾病状态和正常生理状态的蛋白质染色斑点的分布会出现差异,以此为线索,可以发现新的药物靶标。

质谱分析(MS)技术具有高通量、敏感性强的特点,能根据相关序列识别蛋白质。其主要作用是识别不同样品中大量相关蛋白质的差异,根据这些差异来筛选可能的疾病相关蛋白,然后与临床实验作比较,以确定真正的靶标蛋白。

在蛋白质组学研究中,进行2-DE和MS研究还需要使用许多其他相关技术。如样品的制备和分离技术、蛋白质结构的分析技术、生物信息学技术等。利用蛋白质组学技术发现药物靶标的一般流程是:样品制备(sample preparation)→分离(fractionation)→质谱分析(mass spectrometry)→蛋白质阵列(protein arrays)→计算生物学(computational biology)→结构蛋白质组学(structural proteomics)→结合特征分析(binding characteristics)^[6]。

另外,酵母双杂交技术也是发现药物靶标的重要途径。该技术能够通过报告基因的表达产物敏感地检测到蛋白质之间相互作用的路径。对于能够引发疾病反应的蛋白互作用,可以采取药物干扰的方法,阻止它们的相互作用以达到治疗疾病的目的。例如,Dengue病毒能引起黄热病、肝炎等疾病,研究发现它的病毒RNA复制与依赖于RNA的RNA聚合酶(NS5)、拓扑异构酶NS3以及细胞核转运受体 β -importin的相互作用有关。如果能找到相应的药物阻断这些蛋白之间的相互作用,就可以阻止RNA病毒的复制,从而达到治疗这种疾病的目的^[7]。

基于靶标分子结构的药物设计指的是利用生物大分子靶标及相应的配体-靶标复合物三维结构的信息设计新药。其基本过程是:①确定药物作用的靶标分子(如蛋白质、核酸等);②对靶标分子进行分离纯化;③确定靶标分子的三维结构,提出一系列假定的配体与靶分子复合物的三维结构;④依据这些结构信息,利用相关的计算机程序和法则进行配体分子设计,模拟出最佳的配体结构模型;⑤合成这些模拟出来的结构,进行活性测试。若对测试结果感到满意,可进入前临床实验研究阶段,反复重复以上过程,直至满意为止^[8]。

基于靶标分子结构的药物设计需要采用X射线衍射分析和核磁共振波谱(NMR)等结构生物学的研究手段,对靶标蛋白质的分子结构进行深入研究,获得相关信息,借助计算机技术建立靶标的蛋白质结构模型。如治疗艾滋病的安瑞那韦(amprenavir agenerase)和奈非那韦(nelfinavir viracept)就是利用人类免疫缺陷病毒(HIV)蛋白酶的晶体结构开发的药物^[9]。

药物筛选重点发展技术主要包括:①针对多靶标的集群式高通量筛选(high throughput screen)技术;②针对致病基因的调控通路或网络,运用系统生物学方法,发展高通量筛选技术。

药物设计重点发展技术主要包括:①发展基于多个蛋白质三维结构的高通量虚拟筛选方法;②发展新的 ADMET(absorption, distribution, metabolism, excretion and toxicity)预测方法;③发展基于疾病基因调控通路或网络的高通量虚拟筛选方法。

执笔:沈月全

讨论与审核:沈月全 闫晓洁

资料提供:闫晓洁

参考文献

1. 高尚荫. 试论分子生物学. 武汉大学学报(理学版), 1983, 3: 90~95
2. Portugal FH, Cohen JS. A Century of DNA. Cambridge: The M. I. T. Press, 1979
3. 高尚荫. 再论分子生物学——科学的革命. 中国病毒学, 1987, 1: 1~2
4. 李思经. 2001 年的分子生物学. 中国生物工程杂志, 1992, 12(4): 50~52
5. 汪世华, 胡开辉. 蛋白质组学的应用与发展趋势. 福建省科协第五届学术年会提高海峡西岸经济区农业综合生产能力分会场论文集, 2005, 122~123
6. Manabe T. Combination of electrophoretic techniques for comprehensive analysis of complex protein systems. Electrophoresis, 2000, 21: 1116
7. 杨齐衡, 李林. 酵母双杂交技术及其在蛋白质组研究中的应用. 生物化学与生物物理学报, 1999, 31(3): 221~225
8. 来鲁华, 徐筱杰, 唐有祺. 蛋白质的结构预测和分子设计. 自然科学进展, 1995, 5(01): 2~6
9. 来鲁华, 骆兆文, 徐筱杰. 基于蛋白质结构知识的合理药物分子设计. 生物化学与生物物理进展, 1993, 20(03): 176~181

第二节 分学科创新方法的发展历程

一、蛋白质组学研究思路的发展历程(方法论的角度)

生命体最重要的组成部分是基因和蛋白质,其中基因是遗传信息的携带者,而蛋白质是生物功能的执行者,它具有自身的活动规律,因此仅从基因的角度来研究生命活动规律是远远不够的,必须研究由基因转录和翻译出蛋白质的生理过程,才能真正揭示生命活动规律。在对蛋白质系统研究的基础上产生了研究细胞内蛋白质组成及其活动规律的新兴学科——蛋白质组学(proteomics)。蛋白质组(proteome)最早是由澳大利亚的 M. R. Wilkins 于 1994 年首先提出,是指一个细胞或组织所表达的全部蛋白质。蛋白质组学是对不同时间和空间发挥功能的特定蛋白质群体进行研究的学科,不仅包括蛋白质的定性和定量,还包括它们的定位、修饰、活性、功能、相互作用等。它从蛋白质水平上探索蛋白质的表达模式及功能模式,从而为药物开发、新陈代谢途径的调节和控制等提供基础和理论依据^[1]。蛋白质组学不同于传统的蛋白质学科,是在生物体或其细胞的整体蛋白质水平上进行的,是从一个机体或一个细胞的蛋白质整体活动来揭示生命的规律。由于蛋白质具有多样性、可变性、复杂性以及表达量低难于检测等问题,对它们的系统研究非常困难。目前对蛋白质组的研究总体可以分为两个方面,即对蛋白质表达模式(蛋白质组成)和蛋白质功能模式(蛋白质相互作用网络关系)的研究。对蛋白质组研究可以提供如下信息,从基因序列预测的基因产物是否以及何时被翻译,基因产物的相对浓度,翻译后被修饰的程度等。

蛋白质组学的研究经历了一个从点到面,从单个的技术分析到综合技术应用的发展历

程。在技术方面,早期主要使用单一的技术,如二维电泳,质谱等对单个的样品进行分析;在研究对象方面,早期主要针对单一材料,如单一的细胞或组织进行研究。经过多年的发展,蛋白质组学已经进入到利用综合的多种技术,对复杂的细胞,组织,甚至个体进行多层面的研究,例如不同的条件下蛋白的表达,蛋白的结构,蛋白的功能,以及蛋白与蛋白的相互作用等多个方面^[2]。

蛋白质组学技术的发展使我们很快能确定疾病(如癌症)的起因,蛋白质研究分析速度的提高使生命科学的研究工作更深入、更准确、更快速,更接近生命的本质。蛋白质组学是研究细胞内所有蛋白质及其动态变化规律的科学,它能从更深层次上去揭示生命活动的规律,从而对人类生活质量的提高和人类寿命的延长起到巨大作用^[3]。

虽然目前生物信息学的核心是基因组信息学,但研究蛋白质序列、蛋白质结构以及蛋白质的功能也是生物信息学的一个重要方面。基因是重要的生命信息,但是,基因需要通过表达将信息转移到蛋白质上,蛋白质才是生命活动的主要承担者。人类基因组计划已经勾画出人体遗传信息的蓝图,然而,要想全面了解复杂的人体,还需要进一步深入认识基因组所产生的全部蛋白质和 RNA,这是蛋白组学研究的推动力。

蛋白质组是指由基因组编码的全部蛋白质。但在生物体内,蛋白质的数目并不等于基因组内对蛋白质进行编码的基因数目。首先,在一个细胞中,并不是所有基因都同时表达,因而,一个细胞的蛋白质组中蛋白质的数目少于基因组中基因的数目。但是,从基因可变剪切和蛋白质修饰的角度看,蛋白质的数目又远远多于基因组中基因的数目。基因组基本上是固定不变的,而蛋白质组是动态的,具有时空多变性和可调节性,能反映基因的表达时间、表达量,以及蛋白质翻译后的加工修饰和亚细胞分布等。蛋白质组学是研究细胞内所有蛋白质及其动态变化规律的科学,旨在阐明生物体全部蛋白质的表达模式及功能模式,其内容包括鉴定蛋白质的表达、存在方式(修饰形式)、结构、功能和相互作用等。从蛋白质组的定义上就可以清楚地看出,蛋白质组学传统的蛋白质研究不同之处在于,它的研究是在生物体或其细胞的整体蛋白质水平上进行的,它从一个机体或一个细胞的蛋白质整体活动的角度来揭示和阐明生命活动的基本规律。

(一) 表达蛋白质组

早期蛋白质组学的研究主要集中在通过二维电泳等方法分析同一细胞和材料在不同条件下蛋白质的表达差异,并从中发现在特定条件下特异表达的蛋白质。经过了多年的发展,目前二维电泳技术仍然是表达蛋白质组学研究的最重要工具之一。早期蛋白质组学的研究范围主要是指蛋白质的表达模式,随着学科发展,蛋白质组学的研究范围也在不断完善和扩充。目前,蛋白质翻译后修饰的研究已成为蛋白质组研究中的重要部分和巨大挑战。蛋白质——蛋白质相互作用的研究也已被纳入蛋白质组学的研究范畴,而蛋白质高级结构的解析即结构生物学,目前也通常纳入蛋白质组学研究范围。蛋白质组的研究实质上是在细胞水平上对蛋白质进行大规模的平行分离和分析,往往要同时处理成千上万种蛋白质。因此,发展高通量、高灵敏度、高准确性的研究技术平台是现在乃至相当一段时间内蛋白质组学研究的主要任务。可以说,蛋白质组学的发展既是技术所推动的也是受技术限制的。蛋白质组学研究成功与否,很大程度上取决于其技术方法水平的高低。蛋白质研究技术远比基因技术复杂和困难,这是因为不仅氨基酸残基的种类远多于核苷酸残基,而且蛋白质有着复杂的翻译后修饰,如磷酸化和糖基化等,给分离和分析蛋白质带来很多困难。

此外,通过表达载体对蛋白质进行的过量表达和大量纯化也不容易,从而使对蛋白质的性质鉴定和分析面临很多问题,这些都对相应的研究技术提出了新的需求和挑战。

蛋白质组研究中的样品制备,通常对细胞或组织中的全蛋白质组分进行分析,也可以进行样品预分级,即采用各种方法将细胞或组织中的全体蛋白质分成几部分,分别进行蛋白质组研究。样品预分级的主要方法包括根据蛋白质的溶解性进行分离,或者根据蛋白质在细胞中不同细胞器的定位进行分级,如专门分离出细胞核、线粒体或高尔基体等细胞器的蛋白质成分。样品预分级不仅可以提高低丰度蛋白质的百分比,还可以针对某一细胞器的蛋白质组进行研究。例如,在临床研究中,对临床组织样本进行分析,寻找疾病的特征性标记,是蛋白质组研究的重要方向之一。但临床样本都是各种细胞或组织的混和物,生理状态也不均一。例如,在肿瘤组织中,发生癌变的往往是上皮细胞,而这类细胞在肿瘤中总是与血管、基质细胞等混杂在一起。所以,常规采用与正常组织进行差异比较的肿瘤,实际上是多种细胞甚至组织的蛋白质组混合物。这就对蛋白质组研究中的样品分离和分析提出了更高的要求。利用蛋白质的等电点和分子量通过双向凝胶电泳将各种蛋白质区分开来是一种很有效的手段,它在蛋白质组分离技术中起到了关键作用。如何提高双向凝胶电泳的分离容量、灵敏度和分辨率以及对蛋白质差异表达的准确检测是目前双向凝胶电泳技术发展的关键问题。目前已发展出与电泳相结合的高灵敏度蛋白质染色技术,如新型的荧光染色技术。质谱技术是目前蛋白质组研究中发展最快,也最具活力和潜力的技术^[4],它通过测定蛋白质的质量来判别蛋白质的种类。当前蛋白质组研究的核心技术就是双向凝胶电泳-质谱技术,即通过双向凝胶电泳将蛋白质分离,然后利用质谱对蛋白质逐一进行鉴定。对于蛋白质鉴定而言,高通量、高灵敏度和高精度是三个关键指标。一般的质谱技术难以将三者合一,而最近发展的质谱技术可以同时达到以上三个要求,从而实现了对蛋白质准确和大规模的鉴定。

蛋白质组研究应用到的分离技术有多种,要想通过一种方法就可以观察到一个组织、一个细胞或一个生物体中的全部蛋白质几乎是不可能的。首先,细胞内蛋白质的表达丰度相差极大,高丰度的蛋白质往往在分离鉴定过程中将低丰度蛋白质掩盖掉,使一些低丰度的蛋白质,特别是调控蛋白质和激酶,很难用目前的方法检测到;其次,蛋白质的表达具有时间和空间上的差异性,这也决定了人们很难获得细胞、组织或生物体的全部蛋白质组成。目前在蛋白质组研究中应用最多的分离技术是二维聚丙烯酰胺凝胶电泳(two dimensional polyacrylamide gel electrophoresis, 2-DE),它也是目前对蛋白质组分辨率最高、重复性最好的分离技术。Smithies 和 Poulik 最早引入了二维电泳技术,他们将纸电泳和淀粉凝胶电泳结合起来分离血清蛋白质。随后,二维电泳技术经历了更多的发展。聚丙烯酰胺介质的引入,特别是等电聚焦(isoelectric focusing)技术的应用,使人们能够基于蛋白质带电属性对其进行分离。目前所应用的二维电泳体系是由 O'Farrell 等于 1975 年首先提出的,其原理是根据蛋白质的两个一级属性,即等电点和相对分子质量的特异性,将蛋白质混合物在电荷和相对分子质量两个方向上进行分离。蛋白质混合物在第一维方向上的分离是利用蛋白质等电点的不同将其在大孔凝胶中离开,这一过程被称作等电聚焦。蛋白质是两性分子,根据环境 pH 的不同分别带正电、负电或零电荷,在 pH 高于其等电点的位置时,蛋白质带负电,反之带正电。在电场作用下,蛋白质分子会分别向正极或负极漂移,当到达与其等电点相同的 pH 位置时,蛋白质不带电,就不再发生漂移。蛋白质混合物在第二维方向上的分离是按照蛋白质相对分子质量的大小进行分离。蛋白质是带电的生物大分子,在第二维

方向按其相对分子质量分离时,为了消除电荷干扰,需要采用 SDS 对蛋白质进行变性处理。SDS 是一种强离子去污剂,作为变性剂与助溶性试剂,可以断裂分子内与分子间的氢键或其他非共价键使分子变性,破坏蛋白质分子的二级结构与三级结构;同时,强还原试剂巯基乙醇和二硫苏糖醇能使半胱氨酸残基之间的二硫键断裂。在样品和凝胶中加入 SDS 和还原剂后,蛋白质分子被解聚成它的多肽链。解聚后的氨基酸侧链与 SDS 充分结合形成带负电荷的蛋白质-SDS 胶束,所带的电荷大大超过了蛋白质分子原有的电荷,这就消除了不同分子之间原有的电荷差异。因此,这种胶束在 SDS-PAGE 中的电泳迁移率不受蛋白质原有电荷的影响,而主要取决于蛋白质或亚基的大小。

二维电泳存在着繁琐、不稳定和低灵敏度等缺点,发展可替代或补充二维电泳的新方法已成为蛋白质组研究技术最主要的目标之一。目前,二维色谱(2D-LC)、二维毛细管电泳(2D-CE)、液相色谱-毛细管电泳(LC-CE)等新型分离技术都有补充和取代二维电泳之势。另一种策略则是以质谱技术为核心,开发质谱鸟枪法(shot-gun)、毛细管电泳-质谱联用(CE-MS)等新策略直接鉴定全蛋白质组的混合酶解产物。随着对大规模蛋白质相互作用研究的重视,发展高通量和高精度的蛋白质相互作用检测技术也被科学家所关注。此外,蛋白质芯片的发展也十分迅速,并已经在临床诊断中得到应用。在基础研究方面,近两年来蛋白质组研究技术已被应用到各种生命科学领域,如细胞生物学、免疫生物学、神经生物学等^[5~6]。在研究对象上,覆盖了原核微生物、真核微生物、植物和动物等范围,涉及各种重要的生物学现象,如信号转导、细胞分化、蛋白质折叠等。在未来的发展中,蛋白质组学的研究领域将更加广泛。在应用研究方面,蛋白质组学将成为寻找疾病分子标记和药物靶标最有效的方法之一。在对癌症、早发性痴呆等人类重大疾病的临床诊断和治疗方面,蛋白质组技术也有十分诱人的前景,目前国际上许多大型药物公司正投入大量的人力和物力进行蛋白质组学方面的应用性研究。在技术发展方面,蛋白质组学的研究方法将出现多种技术并存,各有优势和局限的特点^[7]。除了发展新方法外,蛋白质组学研究更强调各种方法间的整合和互补,以适应不同蛋白质的不同特征。另外,蛋白质组学与其他学科的交叉也将日益显著和重要,特别是蛋白质组学与其他大规模科学如基因组学,生物信息学等领域的交叉,所呈现出的系统生物学研究模式,将成为未来生命科学最令人激动的新前沿^[8]。

(二) 功能蛋白质组

随着蛋白质组学的发展,人们不仅想了解蛋白质在不同条件下的特异表达,还想了解这些蛋白在定位、折叠、修饰等功能上的不同和差异。蛋白质组学一经出现,就有两种研究策略。一种可称为“竭泽法”,即采用高通量的蛋白质组研究技术分析生物体内尽可能多乃至接近所有的蛋白质,这种观点从大规模、系统性的角度来看待蛋白质组学,也更符合蛋白质组学的本质。但是,由于蛋白质表达随空间和时间不断变化,要分析生物体内所有蛋白质是一个难以实现的目标。另一种可称为“功能法”,即研究不同时期细胞蛋白质组成的变化,以发现随不同环境特异表达的蛋白质,这种观点更倾向于把蛋白质组学作为研究生命现象的手段和方法。一般来讲,大部分细胞生命活动发生在蛋白质水平而不是 DNA 水平,因此即使知道了全部基因的表达概况也难以阐明其实际功能,基因在生物体的功能最终由其编码的蛋白质在细胞水平上体现。从基因表达的角度看,蛋白质组中蛋白质的数目总是少于基因组中开放读框(open reading frame, ORF)的数目,但从蛋白质修饰的角度看,蛋白质的数目又远远大于基因组中 ORF 的数目;基因组基本是不变的,而蛋白质组是动态

的,具有时空性和可调节性,能反映某基因的表达时间、表达量,以及蛋白质翻译后的加工修饰和亚细胞的分布等,功能蛋白质组学的概念由此而生,它指在特定的时间、特定环境、和实验条件下基因组活跃表达的蛋白质,功能蛋白质只是总蛋白质组的一部分^[9]。功能蛋白质组学研究是位于对个别蛋白质的传统蛋白质研究和以全部蛋白质为研究对象的蛋白质研究之间的层次,是细胞内与某个功能有关或某种条件下的一群蛋白质。

一百多年来,特别是 20 世纪的生物化学与分子生物学的研究使得我们认识到,蛋白质是一类结构和功能高度多样,并能对环境做出自发响应的、复杂而神奇的生物大分子。自 1938 年这类分子被正式命名为“蛋白质(protein)”以来,人类对其认识的历史是曲折的,期间曾出现过大量的错误理论。即使到今天,我们对那些存在于生物体内的成千上万种功能和结构各异的蛋白质的认识还有许多未知的地方。蛋白质的基本结构单位是 20 种氨基酸,氨基酸之间通过酰胺键共价连接在一起形成多肽链,多肽链进而折叠成具有特定三维空间结构的大分子。只有组装好的蛋白质分子才能发挥其特定的生物学功能。生物体内成千上万种蛋白质分子几乎参与了生命的所有过程^[10]。

血红蛋白可能是人类了解其相应生物学功能最早的蛋白质。在 1864 年,人们利用分光光度计已经观察到了血红蛋白具有可逆结合氧气的的能力。同时,人们也逐渐认识到了氧气参与的细胞中的氧化反应是生物能量产生的重要途径。这两方面知识的结合,使得人们逐渐认识到了血红蛋白的功能是在脊椎动物血液中输送氧气。在 1926~1930 年间,Sumner 和 Northrop 通过结晶和活性测定研究揭示,具有生物催化功能的酶分子原来也是蛋白质,这是对蛋白质功能认识的一次飞跃。到目前为止,已经被正式命名的酶类蛋白已经有 3000 多种。在目前的研究工作中,蛋白质功能的研究可大致分为两种模式:一是先分离鉴定某种蛋白质,如肌球蛋白和肌动蛋白等,然后再试图去揭示其生理功能;另外就是从已知的特定生理功能,如免疫、防御、视觉等开始,去揭示参与的蛋白质种类。这两种研究的模式互相支持和补充,是目前蛋白质功能研究的主要技术方法,特别是对通过基因组测序所推测的大量未知蛋白的功能的认识,将会是一个巨大的挑战。人类关注蛋白质功能很大程度上是由于蛋白质和人类健康之间千丝万缕的联系,目前已发现的大多数遗传性疾病为基因突变所导致的相应蛋白质的功能异常引起。例如,隐性遗传性的苯丙酮尿症是由苯丙氨酸羟化酶的缺乏造成;白化病是由于先天性缺乏酪氨酸酶,或酪氨酸酶活性下降,而使得黑色素合成发生障碍所导致;遗传性囊性纤维化疾病与位于细胞质膜上的氯离子通道调节因子的功能缺失有关。因为蛋白质功能异常可以引起疾病,人们因此试图通过在活细胞中转入能表达正常蛋白质的基因来治疗相关的疾病,不过这样的治疗方式目前还处于试验阶段。

每一种生命现象,即使是单细胞生命,都会涉及众多不同的生理过程。20 世纪生物化学和分子生物学的蓬勃发展,使得我们对生理过程的认识也从早期的整体宏观水平的描述,上升到了目前的细胞和分子水平的分析。在对一些模式生物,如大肠埃希菌、酵母、线虫、果蝇、拟南芥、人类的测序完成后,各组科学家都试图对所预测的编码蛋白质进行功能分类。那些完成基本生物功能的蛋白质可能在不同物种中是高度保守的,但部分蛋白质及其功能却可能是物种特异的。由于其中一半左右的蛋白质的功能还不清楚,使得我们难以对这些研究中提出的蛋白质功能分类的异同进行总结;但科学家已经发现一些涉及遗传物质的复制和转录、蛋白质合成及物质运输、能量及物质代谢、维持细胞形态、与细胞周期调节和信息传递相关,以及与细胞感受内外环境变化等的普遍性蛋白质,存在于所有模式生物中^[11]。模式生物基因组 DNA 序列的测定,以及对所编码的蛋白质分子的预测结果,更

使我们认识到,人们对蛋白质结构和功能的认识还是非常肤浅和不完整的,对蛋白质功能进行研究的方法和系统还远不完善。鉴于目前用于预测基因组 DNA 编码蛋白质情况的电脑程序的不准确性,以及对蛋白质功能认识总体的不完整性,使得这样的数据分析并非十分可靠,所以只能用作大概的参考。

在长期的蛋白质功能研究的历史过程中,人们习惯于首先将待研究的蛋白质对象进行纯化,进而对其结构特征和生物学活性进行体外分析。对血红蛋白、肌球蛋白和大量酶蛋白等功能的认识是这类研究途径的典型例子。这种传统的研究途径仍旧为很多蛋白质研究学者所沿用。但很多时候这样的研究无法有效揭示某种蛋白质在生物体内的功能。这样的蛋白质体外研究,如果是在已知蛋白质体内功能的基础上而开展的话,将会更为有效。在进行体外蛋白质研究的过程中,科学家会尽量去模拟生物体内的条件,但完全的模拟是不可能的,因为生物体及其基本组成单位细胞是个高度有组织的结构,而这才是蛋白质发挥功能的舞台。当蛋白质被纯化后,这些复杂的细胞结构不复存在,很多体内存在的蛋白质与其他蛋白质之间的非共价弱相互作用也随之消失,而且那些在生物体内含量极低的蛋白质也很难被大量地纯化。但随着 20 世纪 70 年代重组 DNA 技术的出现,很多在生物体内含量很低的蛋白质都可以通过先获取其编码基因,然后通过重组表达的方式大量获得。在观察蛋白质在体外条件下表现出来的特征和行为之后,我们必须还要考察蛋白质在体内的特征和行为。在体内,我们需要考察的蛋白质性状很多,如特定蛋白质在生物内的具体位置(比如是在细胞质内、细胞核内、线粒体膜上或基质内、细胞质膜上、还是动态分布于不同的细胞器内等),与其发生相互作用的蛋白质对象及发生相互作用的类型(比如是稳定的还是瞬时的相互作用),与其结合的其他分子(比如是 DNA、RNA、磷脂类、糖类,还是其他小分子有机物等),同源寡聚化蛋白质分子的寡聚状态是否发生改变,蛋白质被共价化学修饰的种类和条件,蛋白质寿命的长短,蛋白质含量的高低等。

有关体内研究蛋白质功能的方法和手段还在不断发展和完善当中。例如,观察蛋白质在细胞内的精确定位可以借助于各种显微镜技术(如共聚焦显微成像技术、免疫电镜技术等);鉴定与特定蛋白质发生相互作用的蛋白质的方法有各种双杂交系统、各类蛋白质片段互补系统、荧光共振能量转移系统、体内化学交联等。目前研究蛋白功能最有力的手段是转基因和基因敲除技术,借助于分子遗传学的途径从改变的表型开始或改变的基因型开始认识蛋白质的功能。这种方法是建立在基因分子生物学的基础之上的。在大约 20 世纪 50 年代,人们认识到,生物体内的所有蛋白质分子的合成信息皆编码于以 DNA 为携带者的基因内,而且这样的信息的表达是通过以 RNA 分子为中介而实现的(即所谓的遗传信息表达从 DNA 到 RNA 再到蛋白质的“中心法则”)。后来的研究发现,存在于细胞内的蛋白质编码基因本身可以被修饰改造,作为中间信息传递者的信使 RNA 分子的结构也可以被加工,从而使得细胞不再产生蛋白质分子;相反,也可以通过基因插入或者调控基因的表达水平使细胞内特定蛋白质的含量提高。通过对以上的“基因型”(即基因组成)被改变的细胞或生物个体的“表型”(即其所表现出来的结构和行为特征)的分析,能够对特定蛋白质的生物学功能有所认识。尽管这样的通过分子遗传学手段改变细胞内特定蛋白质含量的方法,包括“基因敲除”、“RNA 干扰(RNAi)”等目前常用的技术已经趋于成熟,但是利用这样的方法认识蛋白质功能的有效性仍受到一定的限制。究其原因,可能是因为蛋白质的功能都是在特定的时间和空间发挥的,如果观察的条件不对,特定蛋白质的特定功能的减弱或增强可能对细胞或个体的表型并不造成明显可见的变化。例如,有实验室通过 RNAi 技术试图