21 世纪的生物学

后基因组时代中的结构生物学

王大成
(中国科学院生物物理研究所，北京 100101)

2000 年 6 月 26 日，美英两国首脑会同公私两大人基因组测序集团，在华盛顿白宫东厅正式向世人宣告：人基因组的工作草图 (working draft) 已经绘制完成. 以此为标志，在人类跨进 21 世纪历史新纪元之际，生命科学也正在迎来一个崭新的时代，即后基因组时代 (post-genome era). 这是生命科学历史发展中一次新的飞跃. 20 世纪生物学的最基本成就是揭示生命机体的世代遗传主要是以基因为载体的核酸负责，有机体的当代生命活动主要决定于蛋白质的结构与功能，从而将整个生物学推进到以核酸和蛋白质为中心的分子生物学时代. 由此，生命不再是一个谜. 人基因组的全序测定，以密码的形式给出了“生命之谜”的谜底. 但是，迄今我们仍然读不懂这个谜底. 在以基因组全序为基础 (sequence-based) 的后基因组时代中，将有可能从整个基因组及其全套蛋白质产物的结构-功能机理的高度去了解生命活动的全貌，并系统整合有关生物学的全部知识，揭示生命物质世界的各种前所不知的规律，完全揭开生命之谜，进而驾驭生命使之为人类的社会经济生活服务. 这将使生命科学升华到一个新的历史阶段. 站在这双重“世纪之交”的门槛，展望一下结构生物学在开拓生物学新时期中的地位、影响和发展，是很有意义的. 但是，任何中长期的科学预言都是困难的，实为笔者所不能，因此下面谈及的内容大多以已露端倪或已经启动的事件为基础，显示结构生物学跨世纪发展的一些趋势.

1 战略性重要地位

结构生物学是以生命物质的精确空间结构及其运动为基础来阐明生命活动规律和生命现象本质的学科，其核心内容是蛋白质及其复合物、组装体和由此形成的细胞各类组分的三维结构、运动和相互作用，以及他们与正常生物学功能和异常病理现象的关系. 这一学科内涵决定了结构生物学将在后基因组时代中具有战略性的关键地位.

在后基因组时代中，生物学的中心任务是揭示基因组及其所包含的全部基因的功能，并在此基础上阐明有机体的遗传、进化、发育、生长、衰老、死亡等基本生物学问题，以及与人类健康和疾病相关的生物医学问题. 由于基因的功能最终总是通过其表达产物蛋白质来实现的，因此要了解基因的全部功能活动，最终也必须回到蛋白质上来. 如果将有机体视为一极其复杂的现代化工厂，基因组是这一工厂的中央控制室，调控这一工厂正常运转的所有指令都来自这里，其重要性不言而喻；但这个工厂的主要工人是以蛋白质为主体的生物大分子，他们维系着健康机体所需的几乎所有生命活动，如代谢、生长、运动、呼吸、免疫、物质运输、信号传导以至遗传信息的表达和调控. 基因组全序测定使我们掌握了以密码书写的使工厂运转的全部指令，要查明和显示其功能作用显然离不开对执行这些指令的工人——被编码蛋白质的全面和透彻了解. 因此，在人基因组测序之后进一步集中研究蛋白质的结构与功能，是揭示基因组功能的基本途径. 离开蛋白质，基因组所蕴涵的信息库将成为无水之源. 现在知道，以蛋白质为主体的生物大分子的功能主要决定于它们的三维结构或如近期科学文献中常称的“形貌” (shape). 因此，在以功能为中心的后基因组时代中，阐释有机体重要生命活动的规律和生命现象的本质，不仅需要基因组的信息，还必须依靠全面了解相关蛋白质及其复合物、组装体的精细三维结构、运动和相互作用，及其与各种生命活动的功能关系. 事实上，我们毕竟不能只用基因组 DNA 的一维序列去确定生命活动的机理 (mechanism) 与功能途径 (function pathway)，也难于仅用基因知识去阐释疾病发生与发展的分子机理. 例如，20 多年前发现了癌基因，并因此获得诺贝尔奖. 自那时以来，关于这些基因及其相关蛋白质序列已被研究得相当详细了，但其致癌的分子

机理始终不清楚. 直到 1997 年, 第一个癌基因蛋白 Src (酪氨酸激酶 cSrc) 及其同源蛋白 HcK 的精细三维结构测定, 才精确揭示了这种同时具有正常功能又可在一定条件下致癌的“两面派”分子的“变脸”机制. 在此基础上, 开辟了运用分子开/关机理, 寻求使 Src 分子保持关闭状态以使其对细胞无害的防治癌症的新途径. 这一实例充分显示了基因与蛋白质、蛋白质三维结构与生命活动的关系.

显然, 在后基因组时代生物学的发展中, 结构生物学研究将具有战略性关键地位, 是生命科学中最具有挑战性的前沿领域之一.

2 与基因组学“联姻”, 产生新的重大科学工程, 解析细胞中全套蛋白质的结构与功能

尽管在过去的时代中已经对生命活动的许多方面开展了大量深入的研究, 取得了广泛丰富的知识, 但从总体看, 迄今我们还主要是从单个或部分基因及其产物蛋白质的结构与功能来了解生命活动的各个侧面, 处于在分子水平上认识生命现象的分析阶段. 人类基因组测序研究揭示, 人体大约有 8 ~ 10 万编码蛋白质的基因 (在人基因组中, 编码蛋白质的基因的准确数量至今尚未确定, 这里暂且引用这一数字). 这就是说, 从单一的受精卵发育为成熟的个体, 大约需要 8~ 10 万不同的基础蛋白质 (不包括遗传密码翻译后修饰产生的大量蛋白质修饰物), 它们以精确的时间和空间在细胞中产生、活动、消亡, 恰到好处地实现其维持有机体生命活动所必需的功能, 而这种功能的发挥完全依赖于这些蛋白质的三维结构. 因此, 如何获得这些基础蛋白质并阐明它们的精确三维结构及其与生物功能的关系, 展现细胞活动的全景, 进而揭示不同生物间在进化上的本质联系, 以洞察全局的本质认识为基础解决人类在医药、疾病、生产、环境等诸多方面的严峻问题, 是后基因组时代生物学的基本课题, 也是未来世纪结构生物学面临的严峻挑战.

因此, 随着人类基因组全序测定的完成, 阐释其编码蛋白质的结构与功能已成为生命科学的新前沿. 为迎接这一挑战, 结构生物学已开始与基因组学“联姻”, 发展大规模、高批量克隆基因、表达蛋白质、生长晶体、测定蛋白质三维结构的新方法新途径, 以快速获取细胞中基础蛋白质总体的结构-功能图景. 这正在推动一个新的科学领域的诞生, 国际上称之为结构基因组学 (structural genomics), 其主要目标是将由获得细胞中基因总清单开始的生物学革命扩展到获取细胞中所有编码蛋白质的结构与功能总清单, 从而获得对有机体生命活动的全景式认识. 这是一个在规模和影响上都可与人类基因组计划相比拟的新的重大科学工程, 已获得学术界、企业界和政府部门的共同关注和支持. 1998 年已在美国 Argonne 召开了第一届结构基因组学研讨会, 并分别由马里兰基因组研究所 (TIGR)、加州大学国家实验室 (LANA) 和 Burkeley 国家实验室 (LBNL) 启动了测定细菌模式系统中全套蛋白质三维结构的研究计划. 进入 2000 年, 美英等国开始讨论这一新的科学工程中的国际合作问题, 计划在 2000 年 4 月、9 月、11 月连续召开会议, 以讨论和协调国际投资、知识产权、数据释放等方面的方针和政策. 美国 NIH 作为第一期计划已设立 6 个结构基因组中心; 英、德、日、加政府也都建立了各自的研究机构和计划, 第一期投资已超过 1.5 亿美元. 由于结构基因组学研究必然对新药发现、疾病防治产生重大影响, 企业界也行动迅速, 已有 10 家公司推出了自己的研究计划, 如在圣地亚哥的 Structural Genomix 计划在今后几年中解析上千独立蛋白质 (unique protein) 的结构, 这大体相当于过去 40 年所解析的独立蛋白质结构的总和, 以获得与医药和临床相关的信息.

结构基因组学比之于 DNA 测序, 更加复杂和困难. 这是因为蛋白质的获取要比基因复杂得多, 其三维结构的测定至今还很耗时费钱, 国际上测定一个独立蛋白质的平均费用大约是 10 万美元. 显然, 实现结构基因组学的基本目标十分困难, 其最优战略途径尚是有待解决的问题. 目前国际上有下列 4 种主张: (1) 首先解析功能未知的特定蛋白质的结构, 获取其生物功能的线索; (2) 归并那些结构/功能类似的蛋白质为家族, 确认细胞中的蛋白质家族数量并测定每个家族的至少一个有代表性的结构, 从而获得细胞中蛋白质的总体形貌 (general shapes); (3) 选取特定的模式生物 (如病原体细菌), 阐明其中全套蛋白质的结构; (4) 集中研究那些对了解重要疾病和基本生物问题至关重要的蛋白质群的结构与功能.

无论采取何种战略, 结构基因组学的推进都将对未来世纪的生物学产生深刻的影响. 在方法与技术方面, 将产生高产晶体学 (high throughput crystallography), 大规模 NMR 技术 (large scale NMR), 建立快速、批量克隆和表达基因、纯化蛋白质、培养结晶、解析结构的新方法新技术. 科学研究

的观察和体制也将由此发生重大改变,因为在这一综合性的重大科学工程中个人与单一的小组的作用将是十分有限的,科学家必须寻求广泛的合作与学科间的交融,必须在国际水平建立相互协作的新体制. 例如一个以嗜热菌为模式生物的一个结构基因组学先期计划包含了 7 个研究机构的 17 个实验室.

实现结构基因组学的最终目标将使基因组的遗传密码与细胞中蛋白质的精细结构与功能直接联系起来,从而有可能从构成基因的核苷酸 (A, G, C, T) 的线性序列导出相关蛋白质的形貌和功能,最后读懂“生命”这本“天书”. 在这一基础上,将有可能建立一个生物学周期表,它不是由 100 个化学元素组成,而是由 8~ 10 万个结构和功能清晰的基因及其产物蛋白质组成. 由此,我们将在全新的深度和广度整合所有的生物学知识,在总体上重新认识包括人类在内的生物界,形成崭新的生物学观. 回想 16 世纪建立地球的行星地位后形成新的世界观对人类社会和生产的巨大影响,以及 19 世纪化学元素周期表的建立对化学工业和量子力学理论诞生的有力推动,可以预期,结构基因组学最终目标的完成将对下一世纪人类的经济生活和社会生活产生全面而不可限量的影响.

3 实现快速、自动、批量结构测定,复杂结构和动态过程研究将成热点

生命物质精细结构的研究极大地依赖于复杂的技术条件,受限于材料来源,长期以来科学家在这一领域主要是处于“能研究什么就研究什么”的被动状态,研究进程缓慢、耗时、费钱. 在普通的 X 光源中,能用于结构解析的晶体要求有毫米量级的尺寸,能精细解析结构的分子直径大约是 7.5 nm,测定一个新结构需耗时数月乃至数年,平均花费约 10 万美元. 这使得虽然从首次成功测定蛋白质 (肌红蛋白) 的三维结构距今已经 40 年,但已阐明的蛋白质独立结构 (约 1 000~ 1 500) 约为细胞中独立蛋白质数量 (由基因组测序推导为 8~ 10 万) 的 1%~ 2%. 进入新世纪,这种情况将发生根本改变. 随着基因组计划的迅速进展研究,我们将通过表达基因编码蛋白的途径产生人体和自然界中存在的任意蛋白质,提供可自由选择的材料. 这有可能使我们在短时间中获得成千上万的蛋白质用于结构生物学研究. 与此同时,随着新一代同步辐射和高分辨率 (900~ 1 000 MHz) 大规模 NMR 谱仪的广泛应用,生物大分子结构测定方法和技术将发生质的飞跃. 第三代同步辐射产生的 X 射线新光源具有极高的亮度、极细锐的聚焦和可灵活“剪裁”的波长,它与一些新方法的结合将对结构生物学研究产生重大影响. 特别是,近期新磁体技术 (new magnet technology) 的应用使能产生同类光源的设施有可能广泛使用. 整合所有这些新技术新方法,已经开始发展能快速、批量表达基因、纯化蛋白质、培养晶体、收集衍射数据、解析结构,并运用机器人使之自动化的技术系统. 这将极大地降低对用于结构解析的晶体大小的要求 (可小到 μm 量级),增加研究对象的尺寸 (接近 μm 大小),提高结构解析的速度 (可在几天或几星期测定一个新结构),缩短观察结构变化过程的时间尺度到纳秒水平 (图 1).

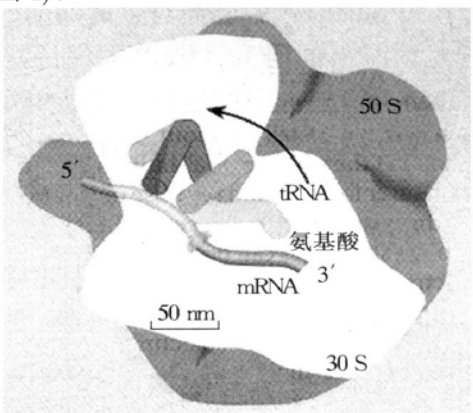

图 1 核糖体——有机体内蛋白质合成的车间

来自细菌的核糖体由 30 S 和 50 S 亚基组成,总共会有 70 多个不相同的蛋白质和 3 条 rRNA 长链分子,是一个极其复杂的组装体. 目前对 30 S 和 50 S 亚基的结构解析已分别达到 0.55 nm 和 0.50 nm,未来 5~ 10 年核糖体的高分辨精细结构将可能获得完全解析.

这些方法和技术的进步使我们有可能测定极大而复杂的蛋白质 (包括膜蛋白)、DNA、RNA 及其复合物和组装体,特别是一些亚细胞器、细胞器原子分辨水平的精细结构,这无疑具有重要的生物学意义. 事实上,几乎所有重要生物学功能都是通过多种重要生物大分子相互作用实现的,因此测定这些大分子复合物、组装体,特别是他们的高级组织形式亚细胞器和细胞器在原子水平的精细结构是结构生物学的重要目标. 但长期以来受方法和技术的限制,这方面的进展很缓慢,事实上在目前已测定结构的上万蛋白质中,复合物仅占 4%,膜蛋白仅有十几个. 在世纪之交,这一方面已经显示良好的发展态势. 应用第三代同步辐射,只用 20~ 40 μm 大小的晶体解析了膜蛋白菌紫红质的结构 (0.25 nm分辨率); K^+ 通道、Cl^- 泵、Ca^{2+} 泵的精

细结构也都在足够高的分辨率解析成功. 包含 8 个组蛋白和一条复杂缠绕的 DNA 链的核小体核心颗粒 (分子质量 206 ku) 的精细结构 (0.28 nm 的分辨率) 在 1997 年底报道, 成为复杂结构研究的里程碑. 最近核糖体 30 S、50 S 亚基和 70 S 完整复合体的低分辨结构也已报道. 今后 5~ 10 年, 更为复杂和重要的核糖体 (76 个蛋白质+ 3 个 rRNA, 2 300 ku), RNA 聚合酶与转录因子的复合物 (60 个蛋白质, 3 500 ku), 核小体、染色质的精细结构都有可能测定. 届时遗传信息复制、转录、翻译的分子机理和翔实过程将获得彻底阐明, 在这一过程中新生肽链的折叠也可能得到全新的了解. 包括膜蛋白在内的复杂结构研究必将突破当前的艰难处境, 打开其精细结构的大门, 成为结构生物学研究的热点. 与此同时, 在第三代同步辐射技术的基础上, 对动态过程研究的时间分辨率已达到纳秒, 纳秒时间分辨晶体学 (nanosecond time-resolved crystallography) 已经应运而生, 并将逐步进入实用阶段. 届时将可以用类似“拍电影”的方法观察和反映生命活动, 使动态过程研究成为新的热点.

结构生物学在未来世纪将发生质的飞跃, 其强大的结构解析能力和深入的研究水平将有可能给出大多数生命物质的精细结构, 使几乎所有重要生命活动的阐释获得坚实的结构基础.

4 查明“构象病”的结构机理, 开辟防治相关疑难病症的新途径

1997 年的诺贝尔生理/医学奖授予了美国人 Prusiner, 因为他发现了一个蛋白质 Prion 是引发曾使世人惊恐的疯牛病和一系列致死性神经变性疾病的病因, 揭示 Prion 分子自身结构的变异 (从以 α 结构为主变为以 β 结构为主) 是疾病产生和传播的基本因素. 由此揭示了一个新的生物-医学原理: 蛋白质可以象细菌、病毒一样是病原体, 它的不适当的三维结构可以引发和传染疾病. 实际上基因表达水平只是决定有多少蛋白质出现在细胞中的一个因素, 基因组 DNA 的一维序列并不能完全表征蛋白质的所有特性, 如生物合成后蛋白质经历的翻译后加工 (如磷酸化、末端氨基修饰等), 从一维氨基酸序列到三维结构的转换 (折叠), 在正确结构指导下的运动和定位, 都不能完全用 DNA 序列描述. 现在知道这些蛋白质被翻译后的每一步如有差错, 都可引发疾病. 近年来已有多种疑难病症被鉴定可能与特定蛋白质的错误三维结构相关, 包括折叠不能引起的疾病 (如囊性纤维化), 有害构象引起的疾病 (如疯牛病和早老性痴呆症), 错误折叠导致错定位引起的疾病 (如高脂血蛋白症), 如表 1 所列. 这是一类“构象病”, 其发生、发展和

表 1 一些可能与蛋白质结构相关的疾病

疾病	突变蛋白/涉及蛋白	类型
有害折叠 (Toxic folds)		
羊瘙痒症/CJD/致死性失眠症/疯牛病	朊病毒蛋白	聚合
Alzheimer's disease	β 淀粉状蛋白	聚合
家族性淀粉样变性	Transthyetin/溶菌酶	聚合
白内障	晶体蛋白	聚合
折叠不能 (Inability to fold)		
囊性纤维化	CFTR	错误折叠/改变 Hsp70 和钙联接蛋白的相互作用
马方综合症 (Marfan syndrome)	肌原纤蛋白	错误折叠
肌萎缩性脊髓侧索硬化	超氧化物歧化酶	错误折叠
坏血病	胶原蛋白	错误折叠
枫糖尿病	α-酮酸脱氢酶复合物	错误折叠
癌症	p53	错误折叠/改变 Hsp70 相互作用
成骨不全症	Ⅰ型原胶原蛋白 Proα	错误组装/改变 Bip 表达
错误折叠导致错定位 (Mislocation owing to misfolding)		
高脂蛋白血 (Ⅱ型) 症	LDL 受体	不当运送
α1-抗胰蛋白酶缺失症	α1-抗胰蛋白酶	不当运送
家族性黑蒙性白痴	β-氨基己糖苷酶	不当运送
视网膜色素瘤 (Retinitis pigmentosa)	视紫红质	不当运送
矮妖精貌综合症 (Leprechaunism)	胰岛素受体	不当运送

传播都直接与蛋白质的三维结构密切相关. 随着人类健康和医疗水平的不断提高, 这类疑难病症的危害性和严重性将会越来越突出, 疯牛病对世界的震撼就是一例. 显然, 了解这类疾病的机理, 寻求有效治疗途径, 离不开相关蛋白质精细结构与功能的研究. 至今, 与这类疾病相关的蛋白质材料的获取、结构鉴定、过程跟踪都极为困难. 以快速、微量、动态为特征的结构解析新方法新技术的应用, 有可能查明如纤维样变性、淀粉样沉淀这类病征的结构机理. 这类疾病相关蛋白质群的结构基因组学研究有可能查明其存在范围、发病和传染机理, 从而提出有效防治途径.

5 批量发现药物靶标, 基于结构的理性药物设计渐成创新药物主流

任何药物都是通过影响人体内的各种生物学过程发挥其治疗作用的, 其主要作用对象是以蛋白质为主体的相关生物大分子, 如各种酶、激素、受体(图 1). 新药的有效设计和开发, 要求对这些药物靶标 (drug target) 的结构、性能有精确的了解. 由于迄今对体内可能的药物靶位了解很少, 现有临床药物绝大多是通过尝试过程 (trial-and-error process) 筛选获得的, 发现一个药物常常需要经历数十年时间, 花费数十亿美元. 近年, 通过对 HIV 病毒蛋白酶的精细结构的测定, 以此为靶设计该酶的抑制剂作为治疗艾滋病的有效药物取得巨大成功, 已经显著减少了艾滋病人死亡数量. 由此显示, 以药物靶分子的精确结构为基础, 有理论依据的 (理性的) 药物设计 (rational drug design) 是发现新药的最有效途径. 但是, 目前按理性方法设计成功的新药仍然很少, 尝试法仍然是新药筛选的主要途径. 随着结构基因组学的展开, 进入下一世纪这种状况将发生重大改变. 目前, 用于临床的蛋白质药物有 59 种, 大多是重组蛋白和单克隆抗体; 已用于临床的药物的分子靶有 483 个, 其分类可见图 2. 有人估计, 通过结构基因组学的基因组编码蛋白质的全结构或结构群体的测定, 有可能发现 800~ 1 000 个新的药物蛋白质 (以基因组编码蛋白质的 1% 计), 提供5 000~ 10 000 个药物靶分子. 在这一基础上, 通过理性药物设计发现的新药将稳定而持续地增长. 这是何等巨大的潜在财富!

正因为如此, 西方大国的企业界正以极大的热情投入结构基因组学的研究, 人们已经开始担心和讨论在政府公共集团与私人企业之间以蛋白质结构为中心可能发生的新一轮激烈竞争. 显然, 在未来世纪, 随着新一代结构生物学的发展, 基于靶分子结构的理性药物设计将可能逐渐成为新药发展的主流.

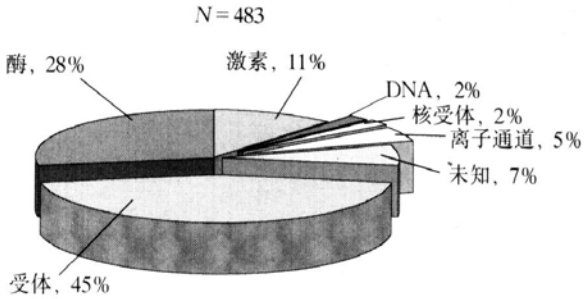

图 2 现有 483 个药物分子靶的分类
这一数字将可能在未来结构基因组学研究中扩大 10~ 20 倍.

参 考 文 献

1 邹承鲁. 结构生物学的时代已经开始. 科技导报, 1995, **11** (4): 7~ 11
Zou C L. Science & Technology Review, 1995, **11** (4): 7~ 11
2 王大成. 结构生物学研究的一些新进展. 生物化学与生物物理进展, 1998, **25** (5): 396~ 403
Wang D C. Prog Biochem Biophys, 1998, **25** (5): 396~ 403
3 王大成. 解析全套细胞蛋白质结构与功能, 展现生命活动全景. In: 李喜先编. 21 世纪 100 个科学难题. 长春: 吉林人民出版社, 1998. 708~ 716
Wang D C. In: Li X X ed. 100 Scientific Puzzles of the 21st Century. Changchun: Jilin Peoples's Press, 1998. 708~ 716
4 Rost B. Marrying structure and genomics. Structure, 1998, **6**: 259~ 263
5 Shapiro L, Lima C D. The argonne structural genomics workshop: Lamaze class for the birth of a new science. Structure, 1998, **6**: 265~ 267
6 Service R. Structural genmics offers high-speed look at proteins. Science, 2000, **287**: 1954~ 1956
7 Service R F. Structural genomics: protein data justice for all. Science, 2000, **288**: 939~ 941
8 Hellemans A. X-ray find new ways to shine. Science, 1997, **227** (5330): 1214~ 1219
9 Service R F. Wiggling and undulating out of an X-ray shortage and the automated approach to protein structure. Science, 1999, **285** (5432): 1342~ 1346
10 Drews J. Drug discovery: a historical perspective. Science, 2000, **287** (5460): 1960~ 1964