结构基因组学研究与核磁共振

刘东升　王金凤*

（中国科学院生物物理研究所，生物大分子国家重点实验室，北京 100101）

摘要　各种生物的基因组 DNA 测序计划的完成，将结构生物学带入了结构基因组学时代. 结构基因组学是对所有基因组产物结构的系统性测定，它运用高通量的选择、表达、纯化以及结构测定和计算分析手段，为基因组的每个蛋白质产物提供实验测定的结构或较好的理论模型，这将加速生命科学各个领域的研究. 生物信息学、基因工程、结构测定技术等的发展为结构基因组学研究提供了保证. 近年来核磁共振在技术方法上的进展，使其成为结构基因组学高通量结构分析中的一个关键方法.

关键词　结构基因组学，基因组，蛋白质家族，三维结构，核磁共振

学科分类号　Q71　A

随着人类基因组计划初步研究结果的发表[1,2]，基因组测序工作已经接近尾声，生命科学正进入一个崭新的时代，即“后基因组时代”，此时，研究的重点变为解读人类 30 亿个碱基对排列顺序所代表的含义，这其中的一个重要问题是要研究它们的表达产物——蛋白质的结构和功能，结构基因组学应运而生了.

结构基因组学是解决全部基因组产物（包括蛋白质和 RNA，鉴于目前进展，本文主要指蛋白质）的三维结构的新兴前沿学科，它综合运用生物信息学、基因工程、X 射线晶体学和核磁共振（NMR）波谱学等的知识，在实测有限蛋白质结构的基础上，通过结构预测阐明基因组中全部蛋白质的三维结构，从此三维结构出发，寻找蛋白质的功能线索. 随着若干个初步研究计划的顺利实施，由美、日等 9 个国家的科研机构组成的“国际结构基因组学科学联合研究体”也已经成立，这标志着以多学科、高通量、新技术、高速度为特征的结构基因组学研究已进入了崭新的阶段.

1　结构基因组学概况

1.1　结构基因组学的研究目标

在传统结构生物学研究中，三维结构往往是生物大分子功能探索中的“终结者”——对一种蛋白质的功能进行透彻研究后，三维结构的研究便理所应当地成为最终揭示结构与功能关系的关键. 随着结构生物学的快速进展，生物大分子的三维结构在生物学中起着越来越重要的作用，从三维结构出发寻找蛋白质的功能，正逐渐成为一个有效的方法[3]，而且，由于“功能”定义的复杂性和不确定性，使用三维结构进行基因组注释虽然并不完美，但至少比功能注释要确定和简单，因此三维结构可能为后基因组时代基因组作“最终”（final）注释[4].

随着人类基因组计划和其他物种基因测序的顺利实施，产生了大量核酸和蛋白质序列. 人类基因组计划初步结果表明，编码人类蛋白质的基因为 3～4 万[1,2]，远低于以前普遍预计的数目. 基因数目少，蛋白质的功能并不会因此而简单. 只有透彻地阐明蛋白质的结构和功能，才能解释人类生命现象为何如此复杂. 截至 2001 年 2 月，PDB（Protein Data Bank）发布的蛋白质三维结构为 1.4 万，它们分别属于一千多个蛋白质家族，因此已知三维结构的蛋白质在整体中只占很少的一部分. 如何快速、大量获得这些基因组蛋白质的三维结构、功能的信息，是摆在人们面前的重要任务.

利用已知的同源蛋白质三维结构数据，由结构预测方法对未知结构蛋白质进行建模，给出较为精确的结果，是大规模获得三维结构的捷径. 当未知结构蛋白质与已知结构蛋白质的序列相似性超过 30％时，由结构预测方法建立的蛋白质三维结构模型已可以用于一般性的功能分析[5]. 这样，有选择地实验测定 10 000～20 000 个靶蛋白质的结构，

* 通讯联系人.

Tel：010-64888490，E-mail：jfw@sun5.ibp.ac.cn

收稿日期：2001-03-09，接受日期：2001-05-17

就可以为所有蛋白质提供同源建模的模板[6].

1.2 结构基因组学研究现状

1998 年 1 月北美 Argonne 会议对结构基因组学研究的诸多方面进行了详细讨论. 在这次会议后，陆续开展了许多初步研究计划. 在 1998～1999 年期间，美国国立卫生研究院（NIH）召开了三次有关蛋白质结构初步行动的研讨会. NIH 资助的七个结构基因组学研究中心于 2000 年 9 月正式运作，参加者主要是根据地区的不同来划分，每个研究中心至少由 5 个研究所组成. 中西部结构基因组中心和东南结构基因组协作组计划从大量不同种类的生物体中挑选新折叠类的靶蛋白质；东北部和纽约两个结构基因组联合体以及结构基因组联合中心重点研究具有重要生理功能的蛋白质家族；柏克利结构基因组中心计划确定生殖器支原体（*Mycoplasma genitalia*）的蛋白质折叠类[7]. 预计在未来几年内，这些研究中心不但可以运用 X 射线晶体学和 NMR 方法测定几百个新蛋白质的结构，还能开发出基因表达、蛋白质纯化、结晶、以及蛋白质结构测定和分析的新方法[8].

日本的 RIKEN 结构基因组初步行动开始于 1995 年，当时主要目标是研究功能域的拓扑结构类型，即折叠类，后来转变为对嗜热栖热菌（*Thermus thermophilus*）中蛋白质家族的结构研究，预期每年将要确定 50～100 个三维结构. 此外，日本的生物信息研究中心（BIRC）将重点研究膜蛋白的结构[9].

欧洲的结构基因组学研究主要集中在高通量结构分析所必需的技术、方法上. Utrecht 等大学正在开发用于结构和功能分析的新方法. York 等大学的研究者们正在为结构分析开发一种一体化的、半自动的和界面友好的软件[10].

目前结构基因组学研究的最大障碍是缺乏必要的协调，这也是大部分制药公司仍在观望的原因之一. 如何使研究交叉最小化而合作最大化，是决定结构基因组学研究发展的重要因素.

2 结构基因组学的研究方面

从技术角度讲，结构基因组学研究大体可分为实验部分和运算部分[11]. 实验部分的工作主要包括：大规模的基因克隆、表达，表达产物的纯化、结晶，X 射线衍射或 NMR 波谱数据收集. 而运算工作主要包括：基因组分析、靶蛋白（家族）的选择、衍射或波谱数据的处理、结构解析以及最后的结构与功能分析. 其流程图如图 1 所示.

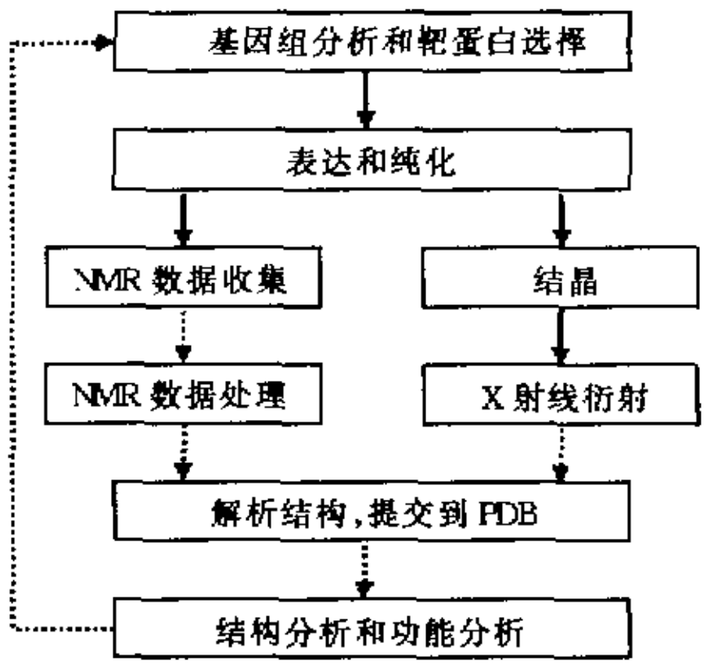

Fig. 1 Main process for the structural genomics

图 1 结构基因组学研究的主要流程

⟶代表物质和信息的转移；⋯⋯▶代表信息的转移.

2.1 靶蛋白质的选择

在传统的结构生物学中，人们往往将目光投向那些功能重要的蛋白质，一些人们普遍关心的蛋白质家族在 PDB 中存有大量（＞20）的结构（图 2a）. 一般来说序列相似性超过 30% 的蛋白质属于同一蛋白质家族. 同源建模的准确度与蛋白质序列相似性有很大关系，通常以序列相似性超过 30% 的实测蛋白质结构为模板，未知结构蛋白质可以被准确建模. 由于 PDB 中存在的上述问题，只有约 5 000 条代表独立的三维结构的记录，可以用于同源建模[11].

结构基因组学研究中选择的靶蛋白的序列应覆盖整个蛋白质序列空间，也就是说，可为每个蛋白质家族提供至少一个实验测定的三维结构，用于整个蛋白质家族的结构建模(图 2b). 根据推测，

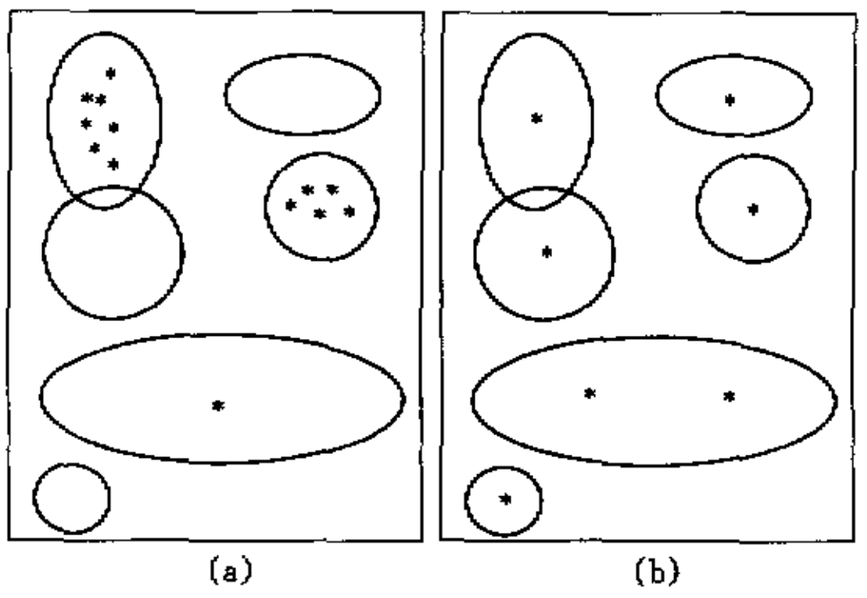

Fig. 2 The difference between traditional structural study (a) and structural genomics (b)

图 2 传统的结构研究(a)和结构基因组学研究(b)的差异

每个圆圈代表一个蛋白质家族，其中的蛋白质序列相似性高于 30%，＊表示三维结构已经由实验确定.

一个独立的结构数据可以提供 15～40 个蛋白质的同源建模[6].

靶蛋白的选择是结构基因组学研究的核心问题，与传统方法有很大不同，它主要通过对数据库的搜索，采用逐步筛选的方法进行. 在选择靶蛋白时，应在覆盖整个基因组范围的前提下使靶蛋白的数量尽可能少，这样可以使实验测定工作更有效，更有代表性. 靶蛋白的选择主要通过三个层次的筛选[12]：生物领域的选择、蛋白质家族的选择和蛋白质的选择.

科学家们在选择生物基因组时有各自不同的考虑，通常选择的生物领域是他们比较感兴趣而且容易研究的生物的基因组. 在 Maryland 大学的 Eisenstein 研究小组选择了流感嗜血菌（*Haemophilus influenzae*）中的 50 种未知结构和功能的蛋白质作为靶蛋白，主要考虑到它的基因组相对较小；突变种的新陈代谢可以操控；含有较少稀有密码子；许多的 ORF（open reading frame）与大肠杆菌相似，便于在其中表达[13]. 日本的 RIKEN 结构基因组初步行动选择了嗜热栖热菌 HB8 的基因组，估计它编码 1 000～2 000种蛋白质，它的好处在于基因组小，蛋白质的热稳定性好而且容易结晶[9].

在研究领域确定后，要排除那些已知结构和能被同源建模的蛋白质家族，包括正在被其他小组研究的蛋白质. 还要排除那些不适合高通量大规模结构基因组学研究的蛋白质家族，主要是排除那些不适合 X 射线衍射和 NMR 研究的蛋白质家族，一般是指那些具有跨膜螺旋序列或者是具有较多低复杂度区域的蛋白质家族. 有些研究组还排除那些有翻译后修饰的蛋白质家族.

在确定所要研究的蛋白质家族后，从蛋白质家族中挑选出一个适合结构研究的蛋白质或蛋白质结构域作为靶蛋白. 主要的根据是蛋白质的大小、热稳定性、等电点等性质.

在确定靶蛋白以后，有些研究小组还为所有的靶蛋白建立一个关系数据库. 其中的信息有：序列信息、表达、纯化、结晶、结构、功能等信息，这些资料可以通过互联网进行查询和交流[13].

2.2　基因的表达与纯化

目前最常用的是大肠杆菌表达体系，一般载体上有 T7 启动子. 为了自动化分离的需要，还要在重组蛋白的 N 端或 C 端加上亲和标签，比如 poly-His，GST-tag 等. 亲和标签的引入可以大大减少纯化步骤，在 NMR 结构测定中，较小的标签（如 poly-His）不必切掉. 据估计，用这样的方法，15%～20%的小蛋白质（不含膜蛋白）可以在大肠杆菌中以可溶形式表达，且能很好地分离纯化. 然而，有约三分之一到二分之一的原核基因并不能很好地表达出适合结晶或 NMR 研究的，具有足够含量和纯度的蛋白质. 而且对于真核基因，这个比例会更高. 此时可以筛选溶解性最好的直系同源基因，还可以更换表达体系，或只表达一个结构域. Edwards 等[14]建立了一个数据库，用来研究蛋白质的表达情况、结晶可能性与序列之间的关系.

2.3　基因组蛋白质的结构测定和结构预测

X 射线衍射和 NMR 是蛋白质结构测定的两大互补方法，在进行三维结构测定之前，要使用质谱、圆二色、一维 NMR、光散射等手段评估样品是否适合三维结构测定. 近年来生物大分子 X 射线晶体学的研究进展快速，给大量确定未知蛋白质的结构创造了良好的条件. 蛋白质结晶一直是 X 射线晶体学中的瓶颈，目前，新型的机器人结晶装置每天可以进行大于 100 000 次的尝试，高通量的蛋白质结晶已经成为可能. 而且蛋白质晶体收集和转移的条件也已大大改善，晶体放置、观测、数据收集等的自动化程度也越来越高. 因此，在结构基因组学研究中，X 射线晶体学仍将承担结构测定的大部分工作. NMR 在结构测定中是 X 射线方法的重要补充，虽然 NMR 研究对蛋白质样品要求较为严格，但是据初步估计，仍然至少有 25% 的酵母 ORF 适合于 NMR 进行结构研究[15].

大部分蛋白质的三维结构可以通过建模来确定，目前建立蛋白质结构模型的方法有三种：从头计算（ab initio）法、比较建模（comparative modeling）法和折叠识别（fold recognition）法[16]. 比较建模法主要是指同源结构预测，是目前蛋白质结构预测中准确度最高的一种. 折叠识别主要是指没有合适的同源模板结构时使用的一种结构预测方法，随着结构基因组学的发展，这种方法可能会逐渐消失[17].

2.4　基因组蛋白质的功能分析

蛋白质的功能是由其特定的空间结构所决定，而蛋白质的空间结构的保守性又远大于序列的保守性，因此在没有同源序列的时候，三维结构将可为蛋白质功能的推测提供大量的信息. 蛋白质的功能至少有三种类型：表型功能、细胞功能和分子功能，而由三维结构揭示的功能一般是指分子水平上的功能或生化功能. 目前存在三种不同的功能分析

类型：由折叠同源性推测功能；由结构特性推断功能；由结构建模推断功能，不同类型的结构分析将提供不同分辨率水平的功能信息[5]. Shapiro 等[18]在研究中发现蛋白质 AdipoQ/Acrp30 的结构与 TNF 家族细胞因子的结构相似，因此推测它的功能与细胞因子相近. Zarembinski 等[19]从高分辨晶体结构发现靶蛋白出人意料地和 ATP 分子结合，因此他们推断这个靶蛋白具有 ATP 酶的活力，生化实验最终证实了他们的推测.

3 结构基因组学中的 NMR

3.1 国际动向

国际上已有许多结构基因组研究中心致力于 NMR 在结构基因组研究中的应用. 如美国东北研究中心计划要解决 NMR 数据的自动分析，而东南研究中心要解决 NMR 结构测定的自动化，目的是要使 NMR 方法适应结构基因组高通量、大规模的结构分析要求. 加拿大多伦多大学组织了结构基因组前沿课题，他们将同位素标记的热自养甲烷杆菌（*Methanobacterium thermoautotrophicum*）基因组编码蛋白质分到几个 NMR 小组，同时进行数据收集和结构分析. 最终在一年内得到了十几个三维结构. 欧洲也已开始着手对高通量 NMR 结构分析计划提供研究基金. 日本的横滨 RIKEN 中心拥有一个庞大的 NMR 实验室. 在实验室中总共安装 10 台 600 MHz，6 台 800 MHz 和 4 台 900 MHz 的 NMR 高场谱仪. 在 2002 年全部安装完毕后，其 NMR 的结构测定数将大体相当于 X 射线的测定数[15].

3.2 结构基因组研究中 NMR 的特点

NMR 在结构基因组研究中是一个关键方法，它的优势体现在以下几个方面：a.NMR 研究不需要蛋白质结晶，而在基因组中有许多蛋白质不能结晶，或者其结晶不能提供很好的衍射图；b.NMR 在接近生理条件的溶液中进行实验，溶液条件稍有不同就会调制蛋白质结构-功能关系，因此可以在不同溶液条件下进行实验，获取有关蛋白质结构与功能的关系；c. 简单的一维同核 NMR 实验，或者相对较容易的二维异核（^{1}H-^{15}N HSQC）实验可直观地提供蛋白质折叠程度（foldedness）的信息；d.NMR 方法所指认的各原子基团的化学位移包含了许多蛋白质功能特征的信息；e. 蛋白质骨架原子核化学位移可以相当正确地确定蛋白质二级结构单元，形成这些二级结构单元的氨基酸片段，对三级结构预测提供了有效的数据；f. 在结构基因组学中，基于结构预测功能的一个重要方面是对下游作用的表征. NMR 波谱中化学位移变化可证实预期的功能，可以筛选配体小分子，扫描配体结合的抗原决定簇.

3.3 NMR 技术进展符合结构基因组学研究的要求

由于 NMR 方法对蛋白质要求较为苛刻，通常要求蛋白质的分子质量不能太大（最大在 25～30 ku），在水溶液中蛋白质要稳定，不降解，不聚合，高溶解度. 在蛋白质表达中，要求高效表达，要求进行同位素标记. 而且具体的 NMR 实验周期长，数据分析所用的机时更长，与结构基因组研究的高通量、大规模的结构分析模式不相符.

近年来 NMR 在脉冲实验方法、数据分析软件以及谱仪硬件技术方面都有长足进展，极大地拓宽了 NMR 的研究范围[15]，也扩展了 NMR 在结构基因组学研究中的应用范围. 首先超低温探头系统研制成功给 NMR 方法注入了新的活力. 由于超低温探头可以极大地提高谱仪的灵敏度（600 MHz 谱仪配置该探头后，其灵敏度将远高于900 MHz谱仪的灵敏度），对稳定性有限，溶解度有限，表达量有限或较高浓度下易聚合蛋白质的研究极为有利. 同时，结合^{2}H 同位素标记蛋白质的制备，还可将 NMR 数据收集时间缩短至原先的 1/10. 现今，800 MHz NMR 谱仪已大量投入使用，900 MHz NMR 谱仪今年已商品化. 在这类高场 NMR 谱仪上运用 TROSY 类型实验还可将 NMR 研究的蛋白质分子质量扩展到 30～35 ku 以上.

蛋白质的 NMR 研究在很大程度上依赖于同位素标记蛋白质样品的制备. 在常用的大肠杆菌表达体系基础上，基于已经成熟的^{13}C、^{15}N、^{2}H 同位素标记方法，近年来发展了许多同位素标记蛋白质的技术方法[20]. 一些单位已经开始运用无细胞表达体系以及细胞系表达体系制备蛋白质样品. 日本 RIKEN 研究中心尝试运用无细胞蛋白质合成系统表达蛋白质，实验发现如果连续向系统补充反应介质，将反应时间延长 10～20 h，则每毫升反应混合物可以提供几个毫克的蛋白质. 这一结果为高通量制备用于结构基因组学研究的蛋白质样品提供了一个途径. 而且在上述的这些表达体系中，可以选择性地标记蛋白质骨架原子核；在^{2}H 均匀标记样品时，可以选择性地质子化甲基基团；通过拼接技术可以只标记蛋白质某一肽段；还可以标记多糖等等. 新技术的发展使 NMR 可用于更多的蛋白质及

蛋白质复合物的研究. 同时，近年开展的在稀释液晶类溶剂中蛋白质残余偶极耦合的研究，更是提供了蛋白质肽键空间取向的信息，强化了 NMR 方法在结构测定中的功能. 期望在不久的将来，可自动化分析 NMR 复杂的大量的数据，那时 NMR 必将在结构基因组研究中作出更多更重要的贡献.

参 考 文 献

1 Venter J C, Adams M D, Mayer E W, *et al*. The sequence of human genome. Science, 2001, **291** (5507): 1304～1351

2 International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. Nature, 2001, **409** (6822): 860～921

3 Shapiro L, Harris T. Finding function throuth structural genomics. Curr Opin Biotech, 2000, **11** (1): 31～35

4 Gerstein M. Integrative database analysis in structural genomics. Nature Struct Biol, 2000, **7** (suppl): 960～963

5 Moult J, Melamud E. From fold to function. Curr Opin in Struct Biol, 2000, **10** (3): 384～389

6 Burley S K. An overview of structural genomics. Nature Struct Biol, 2000, **7** (suppl): 932～934

7 Balasubramanian S, Schneider T, Gerstein M, *et al*. Proteomics of *Mycoplasma genitalium*: identification and characterization of unannotated and atypical proteins in a small model genome. Nuclear Acids Res, 2000, **28** (16): 3075～3082

8 Terwilliger T C. Structural genomics in north america. Nature *Struct Biol*, 2000, **7** (*suppl*): 935～939

9 Yokoyama S, Hirota H, Kigawa T, *et al*. Structural genomics projects in Japan. Nature Struct Biol, 2000, **7** (suppl): 943～945

10 Heinemann U. Structural genomics in Europe: slow start, strong finish?. Nature Struct Biol, 2000, **7** (suppl): 940～942

11 Teichmann S A, Chothia C, Gerstein M. Advances in structural genomics. Curr Opin Struct Biol, 1999, **9** (3): 390～399

12 Brenner S E. Target selection for structural genomics. Nature Struct Biol, 2000, **7** (suppl): 967～969

13 Eisenstein E, Gilliland G L, Herzberg O, *et al*. Biological function *made crystal clear — annotation of hypothetical protein via* strucural genomcis. Curr Opin Biotech, 2000, **11** (1): 25～30

14 Edwards A M, Arrowsmith C H, Christendat D, *et al*. Protein *production: feeding the crystallograpers and NMR spectrocsopists*. Nature Struct Biol, 2000, **7** (suppl): 970～972

15 Montelione G T, Zheng D Y, Huang Y P, *et al*, Protein NMR spectroscopy in structural genomics. Nature Struct Biol, 2000, **7** (suppl): 982～985

16 Jones D T. Protein structure prediction in the postgenomic era. Curr Opin in Struct Biol, 2000, **10** (3): 371～379

17 Murzin A G. Progress in protein structure prediction. Nature Struct Biol, 2001, **8** (2): 110～112

18 Shapiro L, Scherer P E. The crystal structure of a complement-1q family protein suggests an evolutionary link to tumor necrosis factor. Curr Biol, 1998, **8** (6): 335～338

19 Zarembinski T I, Hung L W, Mueller-Dieckmann H J, *et al*. Structure-based assignment of the biochemical function of a hypothetical protein: a test case of structural genomics. Proc Natl Acad Sci USA, 1998, **95** (26): 15189～15193

20 Goto N K, Kay L E. New developments in isotope labeling strategies for protein solution NMR spectroscopy. Curr Opin in Struct Biol, 2000, **10** (5): 585～592

Structural Genomics and Nuclear Magnetic Resonance

LIU Dong-Sheng, WANG Jin-Feng*

(*National Laboratory of Biomacromolecules, Institute of Biophysics, The Chinese Academy of Sciences, Beijing* 100101, China)

Abstract The nearly close of sequencing stages of the various genome projects transforms structural biology into structural genomics. Structural genomics is the systematic determination of all genome products' structures. It uses high-throughput selection, expression, purification, structure determination and computational analysis to provide an experimental structure or a good model for every protein in all completed genomes, that will accelerate scientific study in all areas of biological science. The developments of bioinformatics, gene engineering and structure determination techniques provide the guarantee for structural genomics. Recent developments in the technology of nuclear magnetic resonance make it as a key method of high-throughput structural analysis in structural genomics.

Key words structural genomics, genome, protein family, three dimensional structure, nuclear magnetic resonance

*Corresponding author. Tel: 86-10-64888490, E-mail: jfw@sun5.ibp.ac.cn

Received: March 9, 2001　　Accepted: May 17, 2001