DOI:10.16262/j.cnki.1000-8217.1997.04.010

功能基因组学研究

余 珉 陈雨亭 沈 岩

(中国医学科学院基础医学研究所，医学分子生物学国家重点实验室，北京 100005)

[摘要] 当前人类基因组研究的重心正在由“结构”向“功能”转移。一个以基因组功能研究为主要内容的所谓“后基因组时代”(post-genomics)，即功能基因组学(functional genomics)时代即将到来。功能基因组学主要研究基因的识别及其功能信息的提取、鉴定和开发利用，涉及生物信息学、计算机生物学和生物医学研究等领域。其成果将深化对基因组遗传语言，基因结构与功能的关系，生命的起源与进化，细胞发育、生长、分化，以及疾病发生、发展的分子机理等问题的理解。功能基因组学的研究还将促进生命科学与数学、物理、化学、信息科学、计算机科学以及自动化技术等学科的交叉融汇，刺激相关学科技术领域的发展。

[关键词] 基因组，生物信息，功能基因组学

近两年来，生物基因组大规模测序工作进展迅速。一些模型生物的基因组（如：啤酒酵母）全序列已获得[1]，人类基因组计划（HGP）也有望提前完成，届时HGP将奉献给世界一部蕴含着人体生物信息奥秘的由DNA序列写就的巨著。如何破译这些DNA序列中的基因结构与功能信息，并进一步了解真核细胞中一切生命活动的分子细节，是生命科学界面临的迫切挑战，也是功能基因组学所要解决的问题。人类基因组研究的重心正在由“结构”向“功能”转移，一个以基因组功能信息的提取、鉴定和开发利用为主要内容的“后基因组时代”正在到来[2]。

1 生物信息学与生物信息的同源转换

“生物信息学”(bioinformatics)是在生命科学高速发展的今天诞生的一门新学科，它包括了生物学、计算机科学、数学和统计学等多学科领域的知识。它将各种各样的生物信息，如基因的DNA序列、染色体定位、基因产物的结构和功能，及各生物物种间的进化关系等，进行收集、分类和分析，并实现全生命科学界的信息资源共享。

从病毒、细菌到高等真核生物，其DNA中都有序列相近的同源区，对应在其蛋白质结构与功能上则具有同源性和相似性。这是生物在漫长的进化过程中保持下来的进化的连续性。不同生物间蛋白质结构与功能上的相似性和同源性，使我们可以从已知基因和蛋白的结构与功能，预测未知的同源基因的功能，因此构成了信息的同源转换（homology transfer）。即由实验获得的某生物的蛋白质结构与功能信息可转换到另一生物的未知功能蛋白[3,4]。这就使生物信息学不仅仅是生物信息的收集与分类，更主要是用来指导新的实验研究。

本文于1997年3月24日收到.

应用生物信息理论可将实验获得的某一生物信息置于整个生物进化与发展的过程中去分析，使信息在生物间合理地转换，由已知的生物信息探测未知的生物信息。这种研究方法已成为指导新基因发现及其功能研究的主要路线。要完成信息的同源转换必须拥有信息量丰富的数据库，其中应包含已知基因的核苷酸序列、蛋白质的氨基酸序列、相对应的已知结构与功能特点等，还应有对这些信息进行有机分类的体系，同时需要拥有数据库检索和进行同源转换的系统软件。当一个新的未知功能的基因序列被测出后，应将其与数据库中全部已知基因序列进行比较，寻找同源序列和/或同源基因家族。只要这个基因结构与功能上的同源家族被确认，我们就可以利用这个基因家族中已知基因的结构、功能和它在生物体中的作用，来预测新基因的结构与功能，即进行基因的同源转换。在一些情况下，信息的同源转换可精确地描述新基因的结构与功能。但如果同源区仅限于局部氨基酸序列，同源转换在对这些基因的功能预测上就很有限。在部分实例中，虽然一些基因或蛋白在序列组成上同源性很高，在功能上却发生了强烈的变异。因此，对基因结构与功能的实验研究仍是不可或缺的环节，任何由信息的同源转换所预测的基因和蛋白质的结构功能特点必须经过严格的实验证明才可确认。

应用同源转换把生物信息由已知蛋白的结构与功能特点切换到未知基因，推测未知基因功能，现已成为研究新基因结构功能的一个重要步骤。这种有理论、有系统的推测有助于确立实验的目标与方向，大大减少新基因功能研究的实验过程，做到“有的放矢”。同源转换数据库的信息都来源于实验数据，只有精确的实验结果才可使同源信息的转换更有价值。在啤酒酵母的近 6000 个基因中，65% 已获得一定的功能信息，其中 30% 直接来源于生物学实验，另 35% 来源于同源分析[4-6]。

生物信息学作为一门新兴学科，发展速度很快。各种 DNA 序列库、蛋白序列库、基因结构与功能数据库纷纷建立，各种信息的同源转换方法也日渐完善。信息与实验的完美结合将使基因及蛋白质结构与功能的研究进入一个崭新的阶段。

2 基因完整性的判定、同源分析与蛋白质结构预测

利用计算机进行基因序列分析、比较和对结构与功能预测的最大优势，在于只要知道序列，几乎任何种类的计算机分析都比实验研究要便宜和快捷。在获得一段 DNA 序列后，利用计算机对其进行分析的步骤大致包括：寻找一个完整的编码基因，对基因进行同源关系的检索和对基因产物结构与功能的预测。

首先需通过计算机分析来确定该基因结构的完整性。对于基因编码区的辨认可根据以下几方面的特点：(1) 序列是否具备非编码区的 DNA 序列特点。一些重复序列在蛋白编码区较少见，如果在一段 DNA 序列中含有较多的重复序列，它就不太可能是蛋白编码区（排除法）；(2) 该 DNA 序列与其它已知的编码区序列是否有相似性。利用 Blast X 检索系统将 DNA 序列以 6 种读码框架形式译成蛋白质的氨基酸序列，并在蛋白的氨基酸序列库中进行相似性比较，寻找同源家族；(3) 寻找 DNA 序列中的“碱基出现频率不均一”现象。如外显子剪切位点有一些经常出现的核苷酸排列，外显子内部也存在许多双碱基对出现频率不均一现象，这些特点可帮助确定基因编码区序列；(4) 寻找基因上、下游或内部常见的一些功能位点。如基因编码区上游常见的调控序列 TATA box 等，使用 Promoter Scan 或 Net Gene 等软件系统搜寻序列中这些特有序列位点，可以帮助确定基因的位置及完整性。计算机把所

有这些信息搜集整合在一起，进行综合分析，再经实验验证，最终确定这一 DNA 序列是否含有一个结构完整的新基因[7]。

使用计算机对所获新基因的 DNA 序列进行同源分析，其步骤大致包括：(1) 将该 DNA 序列输入适当的数据库与已知功能和（或）结构的基因序列进行比较，确定新基因中是否含有结构功能上保守的区域；(2) 按同源关系将新基因划归某一基因家族；(3) 根据此基因家族中已知基因成员的结构与功能来预测新基因的功能。

在进行同源检索时，使用不同的软件进行多视角分析比仅使用单一软件更具可靠性。选择不同的检索系统进行数据库检索和同源比较，对检索结果有一定影响。目前常用的三个检索系统：Blast，Fasta 和 Smithwaterman 都可用来作数据库检索，但分别具有不同的检索速度、灵敏度和特异性，应根据不同的检索目的与要求选择相应的检索系统。各种数据库往往为不同的检索目的而建立，使用不同的数据库对检索效果也有较大影响。如拥有高注解率的 PIR 库，其中每个已知的蛋白序列都注有结构与功能信息并分类到超蛋白家族中。检索 PIR 库，找到相应的同源家族，可在很大程度上对此基因的功能进行定位。又如 Prosite 等是对蛋白序列进行分类的数据库，可自动将输入的 DNA 序列转换为蛋白质氨基酸序列进行蛋白质同源检索，大大提高了检索的准确性和灵敏度。而如果要寻找一个新基因，现有的 EST/cDNA 数据库则可提供有用的信息。万维网（WWW）则在各个检索途径和数据库间架设起桥梁，使我们可同时从多个检索途径获得关于新基因的最大信息[8]。

在生物进化过程中，有许多保守的氨基酸序列在蛋白质中经常出现，它们往往有特定的空间结构，但在不同生物的不同蛋白中可能执行不同的功能。一般而言，若某蛋白保守氨基酸区域的高级结构已由实验测定，这一结构信息即可应用于同家族的其它蛋白的同源区域。在蛋白数据库中不仅可以找到蛋白家族成员的氨基酸序列，还有这些蛋白的高级结构特点。因此，可以利用同源检索和计算机分析，预测蛋白质高级结构。若经过检索蛋白质氨基酸序列库得知新蛋白含有某一保守片段时，可通过 Prosite，Block 等系统，寻找相应同源片段的结构模型。从计算机网络中收集组成蛋白的各片段的结构特征信息，使用各种识别蛋白分子折叠区序列特征和结构的分析软件系统，可将整个蛋白的结构信息整合在一起，建立各种排列组合的可能性，寻找合理的分子高级结构模型，最后由实验来验证所建立的结构模型。大量的蛋白序列库和进行结构分析的软件系统都可在 WWW 中找到，并可对预测结果的可信性作出定量评估[4,9,10]。

随着新基因序列、功能和结构的不断发现，每天都有大量的生物信息输入数据库。更新更完善的数据库不断建立，更快更特异的检索手段也不断开发，利用计算机进行 DNA 序列同源分析和蛋白质结构与功能的预测，对生物学的发展产生了很大影响[11]。同源信息检索是一种十分快捷的了解新基因的手段，它使我们从刚刚得到一个基因序列时的一无所知、无从下手，到了解它的同源基因和了解其可能的结构与功能变得十分简单迅速。但并不是对每个序列的检索都会带来有效的信息，一些“孤儿序列”在数据库中没有同源家族，有些序列拥有多个分属不同家族的片段，或由于同源关系的确定方法不同而可归入不同的家族，这都会对同源信息检索结果的判断造成困难。此外，庞大的数据库虽使信息更丰富全面，但也使检索更耗时，一些无关信息干扰同源分类的准确性，使检索的特异性下降。因此，建立更专门的数据库和改进检索方法将是进一步努力的方向。

3 基因功能的实验研究

从同源检索获得的信息最终应该用于指导生物学实验。分子生物学的各种实验方法与技巧仍是研究基因结构与功能的最重要途径。对单个已克隆基因的实验研究技术已有了长时间的摸索与发展,许多技术日臻成熟。但面对生物全基因组功能的破译,需要同时分析成千上万个新基因,原有的一些技术方法则显得过分耗时耗力。最近发展起来的 DNA 微芯片技术,使在全基因组范围内同时进行大量基因功能分析成为可能[12]。DNA 微芯片技术是照相平板印刷术与寡核苷酸化学合成技术的结合,在极小的感光面上利用光来指导已知寡核苷酸序列的合成。利用这种光导化学合成技术已可同时合成 65 000 种不同的寡核苷酸链。通过分子杂交和荧光检测技术可检测到浓度小于 1 pM 样品的杂交信号。这种 DNA 微芯片技术现已应用于酵母全基因组基因功能分析[13]。采用 PCR 与同源重组技术产生一系列基因缺失突变株。大量缺失突变株富集在一起,共同在选择性培养基上培养。PCR 扩增筛选出的突变株分子标记,与微芯片上高密度的寡核苷酸集合(与分子标记序列互补的片段)杂交。杂交后,用氩离子束激发荧光标记,并用光电倍增管在 530 nm 测荧光强度,将数据输入计算机进行分析,获得杂交的定量统计结果。由于每个基因都可找到几十个核苷酸组成的特异寡核苷酸标记,由微芯片杂交结果,可得知某一特定基因的表达是它修复了缺失突变产生的影响。这种方法可对不同条件下某修复突变基因的表达进行定量研究,从而对有关基因的功能进行定位。

DNA 微芯片技术还可用于同时对大量 cDNA 和 mRNA 进行定量检测[14]。对生物发育的不同阶段、不同生理状态下有不同表达的基因进行研究,可了解在生物的生长发育和疾病过程中起决定作用基因的功能。用各种方法测定不同发育阶段细胞的 mRNA 丰度,并进行比较,找出在特定发育阶段和特定状态下高表达的基因并确定其功能,是一个重要的研究方向。在这一重要的研究领域中,DNA 微芯片技术以其快速、灵敏及定量的特点而倍受瞩目。

原有的研究基因功能的实验方法也在不断改进。基因敲除、基因在小鼠胚胎细胞的定向导入,在基因的功能研究中都是有效的途径[15]。而酵母双杂交体系[16]和噬菌体显示[17],利用蛋白质的相互作用研究其功能,也取得了许多成果。这些领域的最新进展可在互联网(internet)上查寻。

基因功能的实验研究虽已有了许多方法,但其进展仍很艰难,尤其是对各种生化机制分子过程的了解十分有限。将简单生物的基因结构与功能信息应用于高等复杂生物是一条指导基因功能研究的思路。目前,对各种简单模型生物基因功能的研究已成热点。

4 蛋白质高级结构的实验研究

每种功能蛋白在生物体内都以特定的空间构型来完成生物分子间的相互识别与作用。构成蛋白质氨基酸序列的保守区域往往也形成保守的高级结构。了解蛋白质的空间结构常常可帮助我们识别不同的蛋白家族及了解蛋白的功能。

X-射线晶体衍射是现今最重要的确定生物大分子三维结构的方法。它需要纯净的蛋白产物,并获得性质均一的结晶体。晶体内部,分子在晶格中有规律的排列。当用某一特定波长的射线入射晶格时,其原子中的电子使射线散射,形成晶体的衍射图谱。晶体衍射图谱的分析需要了解结构因子的相位和振幅,振幅可从衍射图谱上直接测定,而相位则需在晶体中引

入重金属原子，根据其在晶格中的相对位置引起振幅的变化间接测定。根据结构因子的相位和振幅可计算出晶体中任何一点的电子密度，从而获得晶体三维结构。X-射线晶体衍射方法对蛋白纯度及结晶技术要求较高，目前对那些不易溶解，不易结晶的蛋白还无法适用[18]。

核磁共振（NMR）是对 X-射线晶体衍射的补充。它无需晶体，可直接测量溶于水的分子结构和分子的动态改变，获得蛋白质分子中结构易变区的结构特点，并可选择蛋白中某一特定区域作检测。NMR 测的是原子核在磁场中的共振光谱。NMR 利用原子核在旋转振动时具有一定的核矩并产生光谱，在瞬间可被外部磁场屏蔽，使光谱发生色散。由化学键连接的几个原子的核矩与共振波可在光谱上整合来判断它们之间的转角。分子量过大的分子其 NMR 共振线难以分辨，因此，NMR 一般只可测分子量小于 15 kd 的分子。若对 N 端和 C 端进行标记，可使测量范围上升到 30 kd，一般相当于蛋白的一至二个结构域[19]。

对蛋白质空间结构的研究可以了解分子间的识别机制和生化反应的原理，使我们可以预测蛋白质可能具有的功能，也可提示抑制剂的结构特点，为药物设计提供思路。

随着生命科学和技术的不断发展，将有更多更好的研究基因功能的方法出现，将有越来越多基因的功能被了解，对于生命及整个世界的认识也会因此而产生飞跃。

参 考 文 献

[1] Goffeau A, Barrell BG, Bussey H, et al. Life with 6 000 genes. Science, 1996, **274**: 546－567.

[2] Editorial. Entering the total－genomic era. Nat Genet, 1997, **15**: 111－112.

[3] Casari G, Andrade MA, Bork P, et al. Challenging times for bioinformatics. Nature, 1994, **376**: 647－648.

[4] Casari G, Daruvar A, Sander C, et al. Bioinformatics and the discovery of gene function. TIG, 1996, **12**: 244－245.

[5] Oliver S. A network approach to the systematic analysis of yeast gene function. TIG, 1996, **12**: 241－242.

[6] Dujon B. The yeast genome project: what did we learn ? TIG, 1996, **12**: 263－270.

[7] Fickett JW. Finding genes by computer: the state of art. TIG, 1996, **12**: 316－320.

[8] Smith RF. Perspectives: sequence data base searching in the era of large-scale genomic sequencing. Genome Research, 1996, **6**: 653－660.

[9] Baron M, Norman DG, Campbell LD. Protein modules. TIBS, 1991, **16**: 13－17.

[10] Hubbard T, Park J, Lahm A, et al. Protein structure prediction: playing the fold. TIBS, 1996, **21**: 279－281.

[11] Gaasterland T, Sensen CW. MAGPIE: automated genome interpretation. TIG, 1996, **12**: 76－78.

[12] Editorial. To affinity... and beyond. Nat Genet, 1996, **14**: 367－370.

[13] Shoemaker DD, Lashkari DA, Morris D, et al. Quantitative phenotypic analysis of yeast deletion mutants using a highly parallel molecular bar-coding strategy. Nat Genet, 1996, **14**: 450－456.

[14] Schena M, Shalon D, Davis RW, et al. Quantitative monitoring of expression patterns with a complementary DNA microarray. Sciencs, 1995, **270**: 467－470.

[15] Jacobson D, Anagnostopoulos A. Internet resources for transgenic or targeted mutation research. TIG, 1996, **12**: 117－118.

[16] Fields S, Song O. A novel genetic system to detect protein-protein interactions. Nature, 1989, **340**: 245－246.

[17] Huse WD, Sastry L, Iverson SA, et al. Generation of a large combinatorial library of the immunoglobulin repertoire in phage lambda. Science, 1989, **246**: 1275－1281.

[18] Acharya KR, Rees AR. X-Ray diffraction of biomolecules. In: Meyers RA ed. Molecular Biology and Biotechnology: a comprehensive desk reference. New York, VCH Pulishers, Inc. 1995, 969－972.

[19] Gaffney BJ, Maguire BC. Nuclear magnetic resonance of biomolecules in solution. In: Meyers RA ed. Molecular Biology and Biotechnology: a comprehensive desk reference. New York, VCH Pulishers, Inc. 1995, 601－604.

FUNCTIONAL GENOMICS AND GENOMIC FUNCTION

Yu Min Chen Yuting ShenYan

(*National Laboratory of Medical Molecular Biology, Institute of Basic Medical Sciences, Chinese Academy of Medical Sciences, Beijing* 100005)

Abstract With the completion of the sequencing of the yeast genome last year and the sequencing of several other genomes on track, what has been termed 'the post-genomic era' looms ever closer, and questions of what this means and where it will lead to become paramount (Editorial, 1997, Nat Genet, 15, 111). Bioinformatics and computational biology are needed for transcriptional analysis. Characterization of potential regulatory elements in genomic DNA remains a difficult task. As the analysis of genomes moves into large-scale sequencing, identification and annotation of biologically relevant features in the sequence become increasingly complex and important. A combination of experimental and theoretical approaches will be brought to bear on these challenges.

·信 息·

中国科协资助自然科学技术类期刊

通过中国科协和全国科学家们的共同努力，财政部 1997 年度已拨出 300 万元专款，用于对自然科学和技术类学术期刊的资助。目前，中国科协制定了具体资助贴补办法，本着择优资助的原则，重点资助那些能够代表我国学术研究水平的、在国内外有较大影响的、由全国性学会主办的自然科学和技术类优秀期刊。资助条件为：

(1) 国内外主要权威性科技文献检索系统收录的优秀期刊；

(2) 基础性学科的重要学术期刊；

(3) 重要的、高水平的工程技术及高科技类学术期刊。

资助强度分为 3 类：被《SCI CDE》收录的期刊每种资助 10 万元；被《SCI SEARCH》和《EI》收录的期刊每种资助 5 万元；未被国际性检索机构收录的科技期刊，由专家评审委员会进行评审择优资助，每种资助 3 万元。这项措施，体现了党和政府对科技期刊的重视和关心，必将激励我国优秀学术期刊进一步提高办刊质量和学术水平，加快与国际接轨的步伐，使更多的期刊能够被国际著名的检索系统所收录，进入文献快速流通渠道，扩大其国际影响，为在国际上提高我国的科技地位做出贡献。

(科学基金杂志部 田中卓 供稿)