综述与专论

基因组后研究策略*

杜光伟 袁建刚 强伯勤

(中国医学科学院 中国协和医科大学 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 由于基因组计划的进展，数据库中积累了越来越多不知道功能的基因序列，分析这些基因的功能将成为基因组计划完成后的主要任务．文章介绍了几种大规模分析基因功能的策略及程序：基因插断、基因表达的系统分析、高密度 cDNA 杂交、蛋白组分析，并讨论了大规模基因功能分析的前景．

关键词 基因组后，基因的功能分析，基因插断，基因表达的系统分析，高密度 cDNA 杂交，蛋白组分析

在过去的几年中，人类基因组计划(human genome project, HGP) 取得了显著的进展，人类及其他模式生物的基因资料在各种数据库中迅速增长[1]．基因组计划的最后成功将为科学家们进一步探索生命的奥秘提供良好的基础．然而，任何一种生物，特别是高等生物生命过程的分子基础是相当复杂的，例如，从一个受精卵到一个成人并维持其成活和健康，需要在特定的时空精确地调节约100 000个基因的活性．解决这一问题的最好方法是分析不同发育阶段不同组织，健康及患病时基因的差异表达．遗憾的是，在过去很长的一段时间内科学家们一次只能分析一个基因，而且操作很复杂（如基因敲除），从而妨碍了鉴定大量基因的功能．由于上述原因，在许多实验室加紧对人类基因组的 30 亿个碱基测序的同时，一些实验室已逐步开始大规模研究基因的表达以及这些基因是如何协同调节整个生物体的活动，开始了所谓的基因组后时代(postgenome era) 的研究[2]．

目前，大规模分析基因功能主要通过以下策略而得以实现：基因插断（gene disruptions）、cDNA 的测序和杂交、蛋白组（proteome）分析．通过这些方法将有可能在短时间内获得大量与基因功能相关的信息．

1 基因插断

模式生物具有许多与人类在序列及功能上保守的基因，因此，对模式生物基因组的研究将为人类基因组研究提供许多信息．在高等模式生物中，果蝇（*Drosophila melanogaster*）积累了最为丰富的研究资料，而且不同于人类，果蝇基因组上的任何开放阅读框架（open reading frame, ORF）都能被突变并在完整的机体上进行详尽的功能分析，因此，果蝇为研究保守基因的功能提供了一个有效的系统．伯克利果蝇基因组计划（the Berkeley *Drosophila* Genome Project, BDGP）在对果蝇基因组定位和测序的同时，开始了一项基因插断计划（BDGP gene-disruption project），以期最大限度地利用果蝇来获取人类基因功能信息[3]．从而克服了过去在基因组上一次只能分析一个基因的局限．BDGP 基因插断计划主要包括以下内容：

首先，用一个经遗传工程改造的 P 转座子插入果蝇基因组的 ORF，收集大量只有一个 P 转座子插入的果蝇品系，这一品系库将为以后的基因插断和突变的鉴定提供极大的便利．以 P 转座子为探针在果蝇唾腺多线染色

* 国家自然科学基金重大项目资助．
收稿日期：1996-04-02，修回日期：1996-07-25

体上原位杂交得到每一转座子插入位置的高分辨细胞遗传图谱. 一旦了解任何基因的细胞遗传学位置, 即可从 BDGP 库中获得 P 转座子插入该基因或其邻近区域的品系, 利用 P 转座子中含有的一个易于识别的眼色标记以及能再次有效地“原位”转座 (大约在 100 kb 内) 的性质还将得到目的基因或其周围区域的突变 (包括染色体微小缺失).

最后, 通过以下两种方法鉴定被插入的基因: 第一个方法是质粒拯救 (plasmid rescue)[4], 克隆并测定转座子插入的 5′ 侧翼序列, 该序列可作为果蝇基因组物理图谱上的序列标志位点 (sequence tag site, STS), 并与特异的开放阅读框架和基因联系起来. 第二个方法是通过果蝇研究数据共享 (research community) 来鉴定基因. 果蝇品系的增强子诱捕 (enhancer trap)[5] 表达模式 (通常反映突变基因的表达模式) 与 BDGP、果蝇数据库 FlyBase (包括每一系谱的染色体插入数据和遗传互补结果) 相结合形成了“果蝇的百科全书 (encyclopaedia of *Drosophlia*)”, 并通过光盘分发. 利用以上信息可以进一步选择特异的基因来研究其在不同组织中表达模式的基础.

目前, BDGP 已收集了 20% ~ 30% 的必需基因被插断的果蝇品系, 在 100 kb 的间隔上将果蝇基因组的遗传学, 细胞遗传学和物理图谱联系起来.

2 cDNA 的测序及杂交

了解基因功能的一个重要步骤是了解其表达模式. 虽然基因表达涉及 mRNA 的稳定性、翻译后修饰、蛋白质降解等因素, 但是真核生物基因表达调控的主要环节是在转录水平, 因此, 特定序列的 mRNA 是否存在以及存在的数量为相应基因的活性提供了一个良好的指标. 由于 mRNA 在体外容易降解, 一般都将其反转录为 cDNA 来操作. 通过 cDNA 获取表达的信息主要采用测序和杂交两种策略.

2.1 测序的方法

为了大规模测定基因的表达, Okubo 等[6] 提出了基因表达的物理图谱的概念, 其主要内容为: 测定 cDNA 3′ 末端的部分序列, 比较各种不同组织类型细胞中 cDNA 的种类和数量即构成基因表达的人体图谱 (body map). 这种图谱能提供两方面的信息: 一是某一组织类型的细胞在不同的发育、分化阶段及不同的生理状态下所表达的 mRNA 的种类和数量; 二是某一基因在所有组织类型的细胞中在不同的发育阶段及不同的生理状态下是否表达和其表达数量. 但是在测定低丰度 cDNA 时, 由于高丰度 cDNA 的干扰, 这一方法显出了它的不足.

Johns Hopkins 大学 Bert Vogelstein 研究组利用测序的方法, 提出了另一种更为聪明的策略来分析基因的表达——基因表达的系统分析 (serial analysis of gene expression, SAGE)[7]. 这一方法的主要依据是: 如果一段 9 bp 长的序列来自于所有被检测基因的同一位置, 则这段序列可以区别262 144个的基因 (4^9). 这段序列不仅可显示某一基因在特异组织中是否表达, 还可作为基因活性的量度, 作者将其称为 SAGE 位标. SAGE 的主要程序为: 用锚定酶 (在大多数转录子上至少有一个切割位点) 和位标酶 (切割位点距不对称识别位点 20 bp 以上的 ⅡS 限制性内切酶) 两种限制性内切酶切割 cDNA 分子的特异位置产生含 9 bp 长的 SAGE 位标, 连接几十个位标, 克隆, 测序. 每一位标是否出现以及出现的频率将代表某一基因是否表达以及表达的水平. 通过比较不同组织的同一发育时期, 同一组织的不同发育时期或健康和病理状态下的 SAGE 位标, 可以进一步克隆与发育或疾病相关的基因.

利用 SAGE 可以在短期内得到丰富的表达信息, 与前面所谈到的直接测定 cDNA 克隆序列的方法相比, 减少了大量的重复测序, 从而大大节省了研究的时间和费用.

2.2 高密度 cDNA 杂交

除了测序的方法外, 杂交的方法在短期内操作大量基因时具有很大的潜力. 由于最近几年技术的进步, 许多高效处理和分析生物材料的仪器的发明, 几个研究组独立提出了用高密

度 cDNA 杂交来系统地分析大量基因的表达模式的方法[8~12]，从而克服了过去由于缺乏有效、准确的定量检测系统而使杂交的方法未能用于基因表达的大规模分析的局限.

几个研究组提出的方法的基本原理和程序都相差不大，只是细节略有不同[8~11]. 我们将其统称为高密度 cDNA 滤膜分析 (high-density cDNA filter analysis, HDCFA)[11]. HDCFA 一般包括以下几个步骤: a. 高密度 cDNA 膜的制备，将含 cDNA (全长或片段) 的质粒或菌落用机器人以高密度点在尼龙膜上; b. 用来源于组织或细胞的总 mRNA 经反转录制备成同位素探针与膜上的 cDNA 杂交; c. 利用生物影像分析系统 (bioimaging analyzer system) 和相应的软件自动定量分析每一个膜上斑点的放射强度. 由于使用了新的检测手段，利用这一方法不仅可以分析细胞中高表达的 mRNA，还可以分析丰度在 0.01% 以下的稀有 mRNA[10,11].

1995 年底，斯坦福大学的 Schena 等[12]提出了另一种杂交的方法——cDNA 微排列 (cDNA microarray)，这一方法与 HDCFA 的主要差别为: a. 杂交的支持物是玻璃制的显微镜载玻片而非尼龙膜; b. 探针用荧光染料标记; c. 直接检测激发的荧光而不需要放射自显影. 由于使用荧光标记，cDNA 微排列还可以将两种不同来源的 mRNA 标记上不同颜色的荧光染料 (如荧光素和丽丝胺)，等比例混匀两种探针后，与排列在同一载玻片上的 cDNA杂交，分别检测两种荧光. 用这一方案可以分析基因的差异表达，从而减少了由于两次单独的杂交所造成的实验误差.

3 蛋白组分析

对细胞内蛋白质的研究兴趣随着从事基因组研究的科学家们提供了越来越多无法鉴定功能的基因而逐渐增长. 由于蛋白质合成之后，通常还要进行磷酸化、糖基化、去除末端氨基酸、蛋白质剪接等翻译后修饰，仅从核酸的序列并不能完全描述一个蛋白质的结构. 然而，尽管分析不同细胞类型中表达的蛋白已将近 20 年，直到 1995 年才由 Wilkins 和 Williams 提出一个与基因组相对应的名词——蛋白组[13]，这是蛋白质分析技术逐步改进和迫切需要进一步了解细胞和基因功能的必然结果.

蛋白组分析的核心技术为蛋白质双向电泳，其主要原理是: 细胞内蛋白质首先根据所带电荷，然后根据分子质量通过两次电泳而得到分离，一般在双向电泳图谱上一次可以得到几千个代表单一蛋白质的斑点. 根据不同组织中的差异表达可以得到相关蛋白的功能信息. 然而直到 80 年代中期，下列问题一直得不到解决: 首先是重复性差; 其次是只能研究细胞内表达量较高的蛋白; 最为糟糕的是，很难辨认电泳胶上的斑点，因而得到的信息很少.

最近，蛋白质研究者采用固相化 pH 梯度 (immobilized pH gradients, IPG) 替代传统的两性电解质并提出了几种用于比较电泳图谱的软件 (如 MELANIE、GUESE2)，改善了蛋白质双向电泳重复性，从而可以比较不同实验室之间的电泳图谱[13,14]. 辨别双向电泳胶上的斑点的关键是了解某些能用于查询蛋白质或核酸数据库的信息，这个问题通过与以下几种技术的结合使用而得以解决: 氨基酸成分、肽片段分析，部分氨基酸序列测定，蛋白质印迹和重组 cDNA 在细胞内的瞬时表达.

1988 年，由于发明了基质辅助的激光吸收离子化 (matrix-assisted laser desorption ionization, MALDI)[15,16]和电喷射粒子化 (electrospray ionization, ESI)[16,17]两种气化和电离蛋白质的方法，电离的蛋白质在真空下根据其质量与电荷之比得以快速、准确鉴定，使质谱仪可以用于蛋白质及其他生物大分子的研究. 随着近几年的完善，利用质谱已能准确、快速地测定蛋白质的分子质量、分析多肽序列、揭示蛋白质的空间结构[16~18]. 此外，质谱分析还具有所需样品少从而允许辨别表达更低的蛋白的优点. 结合相应的高效数据库查询软件和各种 DNA、蛋白质数据库，将双向电泳图谱上的斑点与数据库上的资料迅速对映起

来[13, 14, 19]，大规模、自动化地辨别双向电泳图谱已成为现实.

伴随着基因组研究的进展，蛋白质研究的数据将为已测序的基因提供大量的与亚细胞定位、细胞和组织的分布、产物的修饰等相关的表达信息，从而进一步阐明基因的功能.

4 展 望

大规模分析基因的功能才刚刚开始，无论是从 DNA、cDNA 或蛋白质着手的分析都还各有优缺点，而且上述的方法都还只提供了某一基因的功能线索而非确切的功能. 但是可以预见，在完成现行的以测出人体全部 DNA 序列为主要目标的基因组计划之后，生物学研究仍然继续需要从整个基因组来分析的方法. 下一步的工作就是要进一步完善或发明各种大规模获得基因功能线索和确证基因功能的方法，读懂这些序列. 对基因组功能的研究必将持续下去并显示出良好的前景.

参 考 文 献

1 Collins F. Ahead of schedule and under budget: the genome project passes its fifth birthday. Proc Natl Acad Sci USA, 1995, **92**: 10821~ 10823

2 Nowak R. Entering the postgenome era. Science, 1995, **270**: 368~ 371

3 Spradling A C, Stern D, Kiss L *et al*. Gene disruptions using P transposable elements: an integral component of the *Drosophila* genome project. Proc Natl Acad Sci USA, 1995, **92**: 10824~ 10830

4 Stellar H, Pirrotta V. A transposabable P vecter that confers selectable G418 resistance to *Drosophila* larvae. EMBO J, 1985, **4**: 167~ 171

5 O'Kane C J, Gehring W J. Detection *in situ* of genomic regulatory elements in *Drosophila*. Proc Natl Acad Sci USA, 1987, **84**: 9123~ 9127

6 Okubo K, Hori N, Matoba R *et al*. Large scale cDNA sequencing for analysis of quantitative and qualitative aspects of gene expression. Nature Genet, 1992, **2**: 173~ 179

7 Velculescu V E, Zhang L, Vogelstein B *et al*. Serial analysis of gene expression. Science, 1995, **270**: 484~ 487

8 Gress T M, Hoheisel J D, Lennon G G *et al*. Hybridization fingerprinting of high-density cDNA library arrays with cDNA pools derived from whole tissue. Mamm Genome, 1992, **3**: 609~ 619

9 Auffray C, Behar C, Bois F *et al*. IMAGE: Integration au niveau moleculaire de l'analyse du genome humain et de son expression. C R Acad Sci Paris, 1995, **318**: 263~ 272

10 Nguyen C, Rocha D, Granjeaud S *et al*. Differential gene expression in the murine thymus assayed by quantitative hybridization of arrayed cDNA clone. Genomics, 1995, **29**: 207~ 216

11 Zhao N, Hashida H, Takahashi N *et al*. High-density cDNA filter analysis: a novol approach for large-scale, quantitative analysis of gene expression. Gene, 1995, **156**: 207~ 213

12 Schena M, Shalon D, Davis R W *et al*. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science, 1995, **270**: 467~ 470

13 Kahn P. From genome to proteome: looking at a cell's proteins. Science, 1995, **270**: 369~ 370

14 Huber L A. Mapping cells and sulb-cellular organelles on 2-D gels: 'new tricks for an old horse'. FEBS Letters, 1995, **369**: 122~ 125

15 Karas M, Bachmann D, Bahr U *et al*. Matrix-assisted ultraviolet laser desorption of non-volatile compounds. Int J Mass Spectrum Ion Proc, 1987, **28**: 53~ 68

16 Senko M W, Mclafferty F W. Mass spectrometry of macromolecules: has its time come. Annu Rev Biophys Biomol Sruct, 1994, **23**: 763~ 785

17 Fenn J B, Mann M, Meng C K *et al*. Electrospray ionization for mass spectrometry of large biomolecules. Science, 1989, **246**: 64~ 71

18 Mann M, Wilm M. Electrospray mass spectrometry for protein characterization. TIBS, 1995, **20**: 219~ 224

19 James P, Quadroni M, Carafoli E *et al*. Protein identification in DNA database by peptide mass fingerprinting. Protein Science, 1994, **3**: 1347~ 1350

Strategies for Postgenome Reasearch. DU Guangwei, YUAN Jiangang, QIANG Boqin (*National Laboratry of Medical Molecular Biology, Institute of Basic Medical Sciences, Chinese Academy of Medical Sciences, Peking Union Medical College, Beijing* 100005, China).

Abstract Attributing to the progress in the Genome Project, more and more gene sequences, most of which are not known any information about function, have been accumulated in database. So analysing the function of those genes will become the main goal after the Genome Project is finished. Some strategies and procedures for large-scale functional analyses of genes, such as gene disruptions, serial analysis of gene expression, high density cDNA hybridization and proteome analyses, are introduced. Furthermore, the prospects of large-scale gene functional analyses are also discussed.

Key words postgenome, functional analyses of genes, gene disruptions, SAGE, high density cDNA hybridization, proteome analyses