基因组范围的蛋白质功能研究方法初探

张 玲* 林澄涛 王 恒

(中国医学科学院 中国协和医科大学 基础医学研究所, 北京 100005)

摘要 基因组大规模测序的进展使人们获得了大量新基因数据, 对这些数据进行基因组范围的规模化功能分析的新方法应运而生. 对其中的结构域融合分析法、系统进化特征法、簇分析法, 结构分析法, 插入突变法和综合分析法做一简要介绍和初步探讨.

关键词 基因组, 蛋白质功能, 研究方法

学科分类号 Q75

基因组大规模测序工作的突飞猛进, 产生了大量功能未知的候选基因. 于是新的难题摆在生物学家面前: 怎样才能快速、高效地解码未知基因, 发现这些基因所编码的蛋白质产物的功能[1].

以往人们常用于研究蛋白质功能的方法包括: 免疫共沉淀法和免疫交联法 (免疫学方法); 不连锁非互补突变缺失法 (遗传学方法) 和酵母双杂交法 (生物化学方法). 这些方法比较费时费力, 无法满足规模化分析的需要. 针对这一问题, 近十年来人们探索了许多新的研究方法. 如应用计算机技术进行的同源性或结构分析为基础的分析方法; 用基因组范围的新实验技术研究 mRNA 表达及其生物化学功能的方法; 建立基因破坏库研究基因功能的方法. 本文重点介绍其中的结构域融合分析法、系统进化特征法、簇分析法、结构分析法、插入突变法和综合分析法.

1 结构域融合分析法

这是一种应用纯计算机分析技术从基因组序列中识别蛋白质与蛋白质之间的相互作用的方法. 例如: 在甲基因组中有分别编码 A、B 两种蛋白质的基因, 而在乙基因组中发现 A 和 B 的基因融合成一条单链, 这种单链就叫罗塞达碑序列 (rosetta stone sequence), 那么可以认为 A、B 是甲基因组中相互作用的两种蛋白质[2].

通过热力学原理分析发现: 融合可以降低 A 与 B 分离的熵值, 从而降低 A 与 B 结合的自由能, 因此 A 与 B 结构域的融合将提高两基因产物的亲和力, 使 A 对 B 的有效浓度增加. 如果 *E. coli* 中细胞蛋白质的浓度是微摩尔级的, 其融合蛋白的有效浓度就可以高达毫摩尔以上[3].

用这种方法研究蛋白质功能的一个经典例子是: *E. coli* 的亚单位 GyrA、GyrB 分别和部分拓扑异构酶 II 序列相似并在酵母的拓扑异构酶 II 中融合成单链 (rosetta stone sequence). 这提示 GyrA 和 GyrB 在 *E. coli* 中是相互作用的 (图 1).

Rosetta stone sequence 是该方法的理论基础, 是共调节相关序列在选择压力下形成的. 由于真核

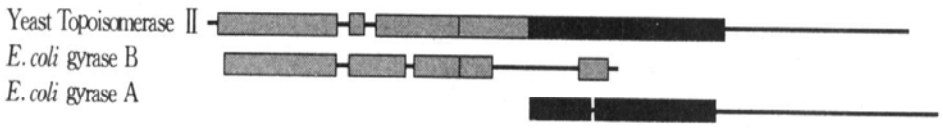

Fig. 1 Rosetta stone sequence[3]

图 1 罗塞达碑序列[3]

gyrase B, gyrase A 分别是 *E. coli* 中两个蛋白质. Topoisomerase II 是酵母中的一个蛋白质. 分析发现 gyrase B, gyrase A 的序列分别和 Topoisomerase II 中的序列同源, 所以认为在 *E. coli* 中这两种蛋白质是相互作用的.

* 通讯联系人. Tel: 010-65296439, E-mail: lingzhy@263.net

收稿日期: 2000-11-13, 接受日期: 2001-01-20

生物基因没有操纵子，不能套用该方法．不过，有研究发现，真核生物相邻基因之间也存在类似的相关作用，如 TCL1 和 Cadherin 蛋白 (Rf)．因而此法对真核基因组的研究可能有一定意义．此方法的另一个缺点是需要在一个基因组中识别出同源序列后，再到另一染色体中寻找相邻序列的基因，因此准确度虽高，覆盖率却很低．用此方法对已验证具有相互作用关系的蛋白质数据库 (ProDom) 进行分析发现，只有 46 对蛋白质由 rosetta stone sequence 偶联，仅占总蛋白质的 4.6%．尽管如此，支持结构域融合分析法的学者认为，随着更多基因组测序的完成，这个比率会逐渐升高．

2 系统进化特征法

图 2 所示的是以 4 个基因组为例说明系统进化特征法的过程．

该方法假设细胞中具有共同作用的蛋白质有共同的进化来源．这一假设是合理的，因为细胞内的蛋白质很少单独起作用．许多代谢途径或复合物在丢失其中的某一成分后作用会减弱．如果一种生物需要某种代谢途径或复合物，通常携带该途径或复合物中的全部基因．反之，不需要这种途径或复合物的生物则完全不含与之相关的任何基因．这种全或无的现象提示了相关功能蛋白质具有共遗传性．通过计算机分析种系发生特征和自动配对，就可以找出共遗传蛋白质．用这种方法曾有人推论 SmpB 蛋白家族对蛋白质合成有作用，这个推论最近被 Karzai 等[5] 证实了．为验证这种方法的可靠性，Matteo 等用两种已知功能的蛋白质，核糖体蛋白 RL7 及鞭毛结构蛋白 FgL 进行研究后证实，两者序列不相似但进化特征相似的特点确实与其功能有关．

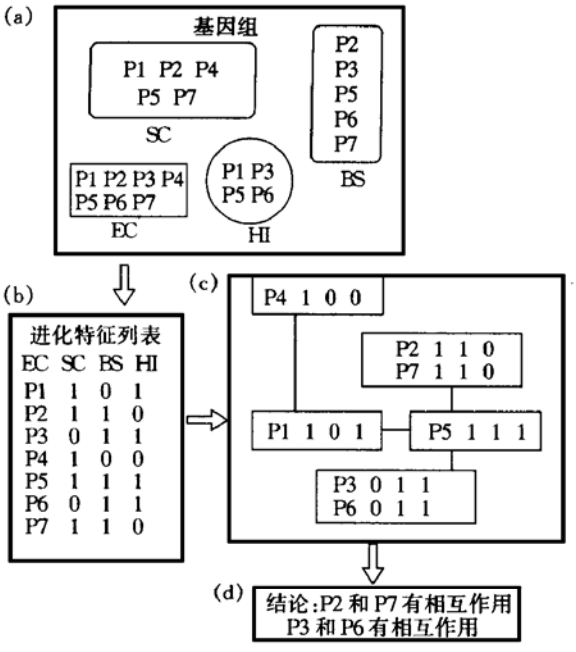

Fig. 2 Protein phylogenetic profiles analysis[4]

图 2 蛋白质系统进化特征分析法[4]

系统进化特征法：P1~ P7 是 *E. coli* (EC) 中 7 个序列不相同的蛋白质．(a) 分别在啤酒酵母 (SC)，流感嗜血杆菌 (HI)，枯草杆菌 (BS) 基因组进行同源性比较；(b) 把比较的结果列表：查到同源性用“0”表示，反之用“1”表示；(c) 按同源比较结果分布规律进行二次列表；(d) 具有相同分布规律的两个蛋白质定义为有相互作用的蛋白质．

实际上，系统进化特征法是与实验遗传学等同的一种计算机方法．把两种生物进行比较时，可以认为一种生物是另外一种生物的基因突变体，是敲除及添加部分基因的集合．把基因按系统进化特征归类可以得到表达图谱 (即同一种生物中基因表达或不表达的情况) 并最终获得标准遗传图．

3 簇 (cluster) 分析法

这是一类跟序列有关的同源分析方法．包括类似性簇分析 (cluster of orthologous group, COG, 直系同源簇方法)，表达簇分析，基因位置簇分析．

3.1 类似性簇分析

这是最大序列相似性搜索方法的改良方法 (进行 COG 分析的站点是 www. ncbi. nlm. nih. gov/COG/)．这种方法是在不同的基因组中寻找类似或相同的蛋白质，定义为对称高分序列，即两个序列在查询时互为最高分序列．由于这种方法不受查询方向的限制，故能有效地把功能类似的蛋白质归类在一起．但这种方法的基础是假设同源= 功能相似．这种理想化的假设决定了它在基因组规模内应用会导致很多错误．鉴于该方法在各种基因及蛋白质的研究中取得了若干成果，因此对于挑选候选基因仍然有价值．

3.2 表达簇分析

该方法通过 DNA 微阵列 (DNA microarray) 对不同生长条件，不同细胞的 mRNA 进行测定，得到表达分析库 (SAGE) 和表达序列标签库 (EST)，然后用统计学方法对表达模式相似性进行归类，并以图形形式表现出来．进行分析的第一步就是用一些敏感的指标 (比如，几何距离，角度，交叉点等) 建立相似性的描述．为更加直观，随后把每个点标以颜色，来定性定量地反映最初的结果．此方法的关键步骤有两个：a. 用一定的算式、

算法从数据中归纳一些遗传规律; b. 通过对比颜色深浅, 找到可视信息所代表的定量信息[6]. 这种方法与系统进化法有些类似, 前提都是蛋白质不单独起作用, 功能相关蛋白质在同时、同位置表达. 只不过研究的对象是不同基因组中的 mRNA.

3.3 基因位置簇分析

这种方法主要用于原核生物蛋白质功能的预测. Demerec 和 Hartman 等早在 1959 年就提出, 基因簇是具有一定位置关系, 且功能相关的一组基因. 不论它们是如何形成的, 自然选择总是阻止其位置的分离. 原核基因簇最明显的特征之一就是它们由功能相关基因组成. 因此人们发明各种算法, 从已知序列计算基因的位置关系, 再按照假设的功能相关基因的位置模式进行筛选. 目前有 3 种因素限制了这种方法的应用: a. 基因以簇形式存在的概率; b. 每个簇的平均大小; c. 含有该基因簇的真实亚系统的大小. 统计学计算表明, 如果两个耦联基因同出现在一个亚系统, 当含该亚系统的基因组数目 ⩾11 时才能发现这对耦联基因. 由于需要多基因组测序数据, 目前染色体基因簇保守性的应用受到限制. 但鉴于在未来的几年里将有上百种基因组被测序, 这种方法将能够从遗传角度描述原核生物的功能耦联[7].

4 结构分析法

对酶的研究发现, 尽管序列不尽相同, 但酶催化活性中心的残基却具有保守性. 于是人们便通过研究蛋白质结构来推测蛋白质功能. 最早的结构研究法是直接寻找蛋白质表面的功能位点, 如较大的表面裂隙或孔穴, 以及分子识别区域. 后来的研究则不仅包括蛋白质序列本身还包括其空间结构. 随着基因组研究计划的进展, 出现了相当多的蛋白质结构数据库. 如 PROSITE (motif 库), DALI 和 ENTREZ (基团距中心的均方根误差), CATH、3Dee (空间结构数据库)[8], PDB, SCOP, HOMSTRAD 等. 结构法的主要缺点是它要求被研究的对象均方根误差 (RMSD) 在 0.4~ 0.5 nm 左右, 由于该预测模型对几何中心分子的描述比环袢周围分子准确, 所以对活性位点在环袢上的蛋白质功能预测就容易出错. 此外, 该方法只能用于酶活性位点研究, 而配基-配体结合模型尚不完善. 尽管结构的相似仅仅是功能相似的条件之一, 但它毕竟是研究蛋白质功能的重要线索, 规模化预测分析需要建立这类数据库.

5 其他研究方法

5.1 插入突变法

该方法是在以往化学突变或放射性突变的基础上建立起来的. 是利用可转座元件插入序列中, 造成一系列突变, 再研究突变后表型变化, 从而了解被突变基因相应的蛋白质功能. 这种经过改良的方法克服了传统方法耗时费力的缺点, 使大规模应用得以实现. 具体可以分为随机插入法和目的插入法[9]. 最近, 用随机突变以及目的突变创建酵母基因破坏库的计划已接近完成 (http://ygac.med.yale.edu)[10]. 除了对酵母的研究[11, 12] 也有人用该方法对细菌[13]、支原体[14]、真菌[15, 16]、植物[17]、线虫[18] 和小鼠[19] 进行类似研究. 这是一种很有应用前景的蛋白质功能研究方法.

5.2 综合计算法

此法是对上述多种方法的综合运用, 旨在建立一种生物蛋白质之间功能联系的网络. Edward 等结合进化相关信息、mRNA 表达模式和结构域融合模式法, 研究了啤酒酵母两种已知功能的基因 Sup35 和 MSH6 (与人类结-直肠癌相关基因同源的酵母基因) 不仅证实了他们已有的功能, 又发现了一些新的功能联系[20]. 可见, 多种方法的综合运用不仅可以提高蛋白质功能预测的准确性, 还可以加快发现新蛋白质功能的进程. 但这种方法得到的联系将相当复杂, 不易分析. 从实验证实的 727 种蛋白质的相互作用中发现了 542 种功能联系, 可见, 即使全部由实验数据组成的网络也是复杂而庞大的. 不过, 这些复杂功能联系的存在恰恰证实了蛋白质很少单独作用, 往往通过物理作用 (physical interaction, 指与分子结构如 motif, domain 等直接相关的相互作用) 或功能联系与其他蛋白质形成作用网络.

综上所述, 这些研究蛋白质功能方法都利用了功能相关蛋白质之间的某些共同点, 包括结构域融合模式、进化共遗传性、相关基因位置的保守性以及表达模式的相似性. 它们的特点是快速、高效和规模化. 纳入这一研究体系的蛋白质大多具有一定的背景, 如位于某种信号通路或复合物中. 因此, 利用某通路或复合物已知功能蛋白质就能把功能研究扩展到其他未知蛋白质. 但是, 除了基因破坏库方法外, 其余方法得到的结果尚需进一步实验验证. 而且, 这些方法所应用的模式主要适用于原核生物, 对于真核生物目前还没有可以进行规模化应用的计算模式. 此外, 很多方法需要多基因组数

据，由于受数据来源的限制而不能广泛适用．尽管如此，上述方法在用理论计算替代繁琐实验方面做出了有益的尝试，并已经取得可喜进展．可以预见，随着数据的完善、新的更有效分析方法的产生以及各种方法的合理综合运用，基因组时代的蛋白质功能研究将向网络化和规模化飞速发展．

参 考 文 献

1 Marcotte E M. Computational genetics: finding protein function by nonhomology methods. Curr Opin Struct Biol, 2000, **10** (3): 359 ~ 365

2 Enright A J, Iliopoulos I, Kyrpides N C, *et al*. Protein interaction maps for complete genomes based on gene fusion events [see comments]. Nature, 1999, **402** (6757): 86~ 90

3 Marcotte E M, Pellegrini M, Ng H L, *et al*. Detecting protein function and protein-protein interactions from genome sequences. Science, 1999, **285** (5428): 751~ 753

4 Pellegrini M, Marcotte E M, Thompson M J, *et al*. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc Natl Acad Sci USA, 1999, **96** (8): 4285~ 4288

5 Karzai A W, Susskind M M, Sauer R T. SmpB, a unique RNA-binding protein essential for the peptide-tagging activity of SsrA (tmRNA). Embo J, 1999, **18** (13): 3793~ 3799

6 Eisen M B, Spellman P T, Brown P O, *et al*. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci USA, 1998, **95** (25): 14863~ 14868

7 Overbeek R, Fonstein M, D Souza M. The use of gene clusters to infer functional coupling. Proc Natl Acad Sci USA, 1999, **96** (6): 2896~ 2901

8 Orengo C A, Todd A E, Thornton J M. From protein structure to function. Curr Opin Struct Biol, 1999, **9** (3): 374~ 382

9 Otto S P. Unravelling gene interactions [news; comment]. Nature, 1997, **390** (6658): 343

10 Coelho S P, Kumar A, Snyder M. Genome-wide mutant collections: toolboxes for functional genomics. Curr Opin Microbiol, 2000, **3** (3): 309~ 315

11 Smith V, Chou K N, Lashkari D, *et al*. Functional analysis of the genes of yeast chromosome V by genetic footprinting. Science, 1996, **274** (5295): 2069~ 2074

12 Smith V, Botstein D, Brown P O. Genetic footprinting: a genomic strategy for determining a gene's function given its sequence. Proc Natl Acad Sci USA, 1995, **92** (14): 6479~ 6483

13 Hensel M, Holden D W. Molecular genetic approaches for the study of virulence in both pathogenic bacteria and fungi. Microbiology, 1996, **142** (Pt 5): 1049~ 1058

14 Hutchison C A, Peterson S N, Gill S R, *et al*. Global transposon mutagenesis and a minimal *Mycoplasma* genome [see comments]. Science, 1999, **286** (5447): 2165~ 2169

15 Burns N, Grimwade B, Ross-Macdonald P B, *et al*. Large-scale analysis of gene expression, protein localization, and gene disruption in *Saccharomyces cerevisiae*. Genes Dev, 1994, **8** (9): 1087~ 1105

16 Kempken F, Kuck U. Transposons in filamentous fungi-facts and perspectives. Bioessays, 1998, **20** (8): 652~ 659

17 Martienssen R A. Functional genomics: probing plant gene function and expression with transposons. Proc Natl Acad Sci USA, 1998, **95** (5): 2021~ 2026

18 Liu L X, Spoerke J M, Mulligan E L, *et al*. High-throughput isolation of *Caenorhabditis elegans* deletion mutants. Genome Res, 1999, **9** (9): 859~ 867

19 Zambrowicz B P, Friedrich G A, Buxton E C, *et al*. Disruption and sequence identification of 2 000 genes in mouse embryonic stem cells. Nature, 1998, **392** (6676): 608~ 611

20 Marcotte E M, Pellegrini M, Thompson M J, *et al*. A combined algorithm for genome-wide prediction of protein function [see comments]. Nature, 1999, **402** (6757): 83~ 86

Primary Research on Genome-wide Protein Function

ZHANG Ling*, LIN Cheng-Tao, WANG Heng

(Etiology Department of Basic Medical Science Institute, Chinese Academy of Medical Science & Peking Union Medical College, Beijing 100005, China)

Abstract Large amount of genome data has been obtained from the rapid progress of genome sequencing. Along with this trend, many new approaches for the study of protein function have been invented. A brief introduction and discussion of those methods such as domain fusion analysis, protein phylogenetic profiles, cluster analysis, protein structure analysis, insertional mutagenesis and the combined algorithm for genome-wide prediction of protein function is reviewed.

Key words genome, protein function, method

* Corresponding author. Tel: 86-10-65296439, E-mail: lingzhy@263.net

Received: November 13, 2000 Accepted: January 20, 2001