植物基因组表达序列标签 (EST) 计划研究进展*

骆 蒙** 贾继增

(中国农业科学院品种资源研究所, 农业部作物种质资源与生物技术重点实验室, 北京 100081)

摘要 植物表达序列标签 (EST) 计划是随机挑选 cDNA 克隆, 并对其 3′或 5′端进行大规模一次性测序, 将得到的 150~ 500 bp 长度的 DNA 片段与数据库中的序列进行比较, 获得对基因组结构、组织、表达等认识的基因组研究策略. 就近年来国际植物 EST 计划的实施情况、植物 EST 计划的研究范围、生物信息学在 EST 研究中的应用、EST 数据库及查询、植物 EST 研究中遇到的问题等方面内容进行了综述.

关键词 植物基因组, 表达序列标签, 生物信息学

学科分类号 Q75

人类基因组计划的提出与实施, 使人类对生命奥秘的探索迈向了一个新纪元. 与人类基因组研究一样, 植物基因组的研究也有两个层次: 一是对全基因组 DNA 序列进行测序, 不仅研究编码基因的序列、组织、物理位置与功能, 而且还要了解重复序列、非转录序列等的大小、组织及功能; 另一个是从表达序列水平的研究, 这就是表达序列标签 (expressed sequence tags, EST) 计划.

1 植物 EST 计划

1.1 EST 计划

EST 是长约 150~ 500 bp 的基因表达序列片段. EST 技术是将 mRNA 反转录成 cDNA 并克隆到载体构建成 cDNA 文库后, 大规模随机挑选 cDNA 克隆, 对其 5′或 3′端进行一步法测序, 所获序列与基因数据库已知序列比较, 从而获得对生物体生长发育、繁殖分化、遗传变异、衰老死亡等一系列生命过程认识的技术[1].

用 EST 技术来进行基因组研究的思想, 是由美国科学家 Venter 等在人类基因组计划开始时提出的, 称为 EST 计划[2]. 这一基因组研究策略的优越性显而易见, 因为表达基因只占整个基因组的 3%~ 5%, EST 反映的是基因的编码部分, 所以 EST 计划可以直接获得基因表达的信息. 另一方面, 用 EST 代替基因组测序, 可使研究费用大大降低, 同时效率大大提高, 具有多、快、好、省之特点. 但 EST 计划也有其不足之处, 主要为所获基因组信息不全, 如调控序列、内含子等在基因表达调控中起重要作用的信息不能体现出来.

1.2 植物基因组计划

在人类基因组计划开始的同时, 一些模式生物的基因组计划也相继开展, 如小鼠、果蝇、线虫等. 在植物中, 拟南芥由于其基因组小、生活周期短、繁殖速度快、易获得较多突变体等特点, 被选为第一个开展基因组计划的双子叶植物. 该计划于 1991 年底由美国、日本、欧盟合作开展, 已有 2.1 万~ 2.4 万个表达基因, 已通过 EST 被鉴定出来, 约占拟南芥基因总数的 70% 以上[3]. 水稻是第二个被选中进行基因组研究的单子叶植物, 这一方面是由于它为基因组最小 (430 Mb) 的重要的粮食作物; 另一方面对谷物的比较基因组学研究表明, 水稻、小麦、玉米、高粱等的基因组成、基因顺序等存在高度共线性, 因此利用对水稻基因组研究的结果, 可从其他谷物中分离与鉴定出对应的基因. 该计划于 1991 年启动[4], 由美国、日本、中国、韩国等国参加, EST 计划为其中重要的组成部分. 国际小麦族 EST 计划 (International Triteace EST Cooperation, ITEC) 于 1998 年启动, 共有 35 个国家与组织参加, 我国也是成员之一 (http://wheat.pw.usda.gov/genome/). 该计划的第一期工作已于 2000 年 7 月结束, 完成 EST 近 3 万条. 此外, 也有许多由各国独立开展的基因组计划, 主要植物包括玉米、小麦、大麦、大豆、棉花、油菜、甘蔗等. 除了这些以农作物为主的草本植物以外,

* 国家 973 计划 (973-08-02) 和国家自然科学基金资助 (399800029) 项目.

** 通讯联系人.

Tel: 010-62186652, E-mail: lmzy@public2.east.net.cn

收稿日期: 2000-09-27, 接受日期: 2000-11-03

近年来也开展了一些木本植物的 EST 研究，如柑橘[5]、杨树[6]、松树[7]等等.

纵观已开展的植物基因组计划，除了拟南芥与水稻在进行 EST 测序的同时，还进行全基因组测序外，其他植物基因组计划实际都为 EST 计划. 虽然 GenBank EST 子数据库中已报道的植物有近百种，但真正称得上 EST 计划的，大多是重要的粮食作物、经济作物和几种特定研究方向的树木.

2 EST 计划的主要研究内容

2.1 构建遗传学图谱

遗传图谱、物理图谱和转录图谱是基因组计划要获得的三张遗传学图谱. 构建染色体物理图谱需要大量的单拷贝短序列 (sequence tagged site, STS) 作为界标，由于大多数基因是单拷贝的，所以 EST 可以用来充当 STS[8]. 同样 EST 片段由于其多态性高可以作为分子标记，用来建立遗传连锁图谱[9]. 如有人用日本的水稻品种 Nipponbare 和印度品种 Kasalath 杂交 F_2 代 186 个植株作遗传图谱，选用了2 300个 DNA 标记，12 个连锁群，总的遗传距离为1 550 cM，其中 70% 的标记来源于 EST[4]. 转录图为染色体 DNA 的某一区段内，所有可转录序列的分布图，EST 为转录基因的产物，可直接用来构建该图[1]，它可以与基因组文库序列比较，提供内含子结构、可选择的剪切方式、转录起始与终止位点等信息.

2.2 分离与鉴定新基因

分离基因的经典方法为图位克隆和转座子标签克隆，是利用分子标记或表型变化来分析鉴定基因. 以 EST 为重要来源的染色体物理图谱进一步方便了对侯选基因的连锁分析，它可将基因确定在更狭小的染色体区段内，缩小了基因的筛选范围. 用 EST 取代对 cDNA 全长的筛选、基因组序列的鉴定等繁琐的实验操作，可大大地提高分离基因的效率. 将所获 EST 用生物信息学方法与各公共数据库中已知序列进行比较，可迅速而准确地确定基因功能. 由于在构建 cDNA 文库时要尽可能地选用全长 cDNA，所以一旦发现有价值 EST，可以找到对应的克隆，获得的全长 cDNA 可以直接用于如转基因等的研究. 利用 EST 方法进行发现、分离基因的研究，不仅是人类基因组研究的热点，而且是植物基因组研究的重要内容[3,4].

2.3 基因差异表达的研究

一般认为某一时期的基因表达数量通常占全部基因的 15%，细胞的分化由基因特异性的时空表达决定. 近年来，对基因的差异表达研究发展了如差减杂交、mRNA 差别显示等新技术，这些技术各有其优势，而 EST 技术的优势在于其稳定性高和分析规模大. 对 cDNA 文库随机挑选克隆进行大规模测序，可直接回答特定组织细胞在某一时期哪些基因表达了，丰度如何等问题，从而能在基因整体水平研究相关的功能及代谢. 目前基因差异表达研究是植物 EST 研究的主流，基因表达谱的研究更是其中的热点，如水稻胚乳发育中的基因表达[10]、拟南芥防御系统基因表达[11]、油菜保卫细胞的代谢研究[12]、杨树和松树木质形成中的基因表达[6,7]等等，涉及的植物种类多，研究内容广泛.

2.4 比较基因组学研究

植物比较基因组学研究是从比较不同植物的遗传图谱开始的，如禾本科植物遗传作图的共线性研究. 利用 EST 中古老保守序列 (ancient conserved region, ACR) 来研究物种之间的进化关系，可以避免选择家系、构建群体、大量的统计分析等周期长而繁琐的遗传图谱制作过程，并且使结果更加准确. 如对拟南芥的研究表明，约 2/3 以上的 EST 中有 ACR 序列，据此可以判定进化关系[13]. 此外利用基因组较小的模式生物所获已知功能基因，用复杂基因组生物如人与作物的 EST 比较，从而推测待测 EST 的功能，亦为比较基因组学研究的重要内容.

2.5 用于制备 DNA 芯片

DNA 芯片技术是近年来发展起来的研究功能基因组学的方法，EST 是用于制备 DNA 芯片很好的基因资源[14]，而芯片技术同样也是目前高通量筛选 EST 的有效方法. 随着更多基因组计划的开展和更多已知功能基因的累积，将来无需测序只通过 DNA 芯片技术就可研究基因功能. EST 计划的实施，不仅获得了关于表达基因的信息资源，同时也获得了大量已鉴定了的 cDNA 克隆资源，而这些资源都是功能基因组学研究不可缺少的.

3 生物信息学在 EST 计划中的作用

3.1 生物信息学

生物信息学 (bioinformatics) 的兴起与人类基因组计划的实施密切相关，它的研究范围覆盖了生物信息的获取、处理、存储、共享、分析和解释等方面. 它综合运用数学、计算机科学和生物科学等知识来阐明各类数据的生物学意义，数据库、计算

机网络、分析软件是该学科的运作平台. 随着基因组计划的实施及其他分子生物学研究的深入, 每天有大量的核酸及蛋白质序列产生, 据统计 GenBank 中核苷酸数每 14 个月翻一番[14]. 面对这些数量庞大且又高度复杂的生物数据, 靠人工完成对其分析与计算是不可想象的, 生物信息学因此而产生, 同时也奠定了其在基因组学研究中的重要地位.

3.2 EST 数据库

数据库是生物信息学的主要内容, 从数据库的种类来看, 核酸和蛋白质序列数据库是最基本的数据库. 目前较为常用的核酸序列数据库有: 美国国家信息中心的 GenBank, 欧洲分子生物学实验室的 EMBL, 日本国家数据库 DDBJ, 这 3 个数据库是收录范围最广并完全向公众开放的数据库, 在它们中均含有 EST 子数据库 dbEST. 在核酸序列数据库中, EST 的量要占 65% 以上[15]. 我国于 1996 年在北京大学建立了生物信息中心 CBI[16], 引进了核酸蛋白质序列、结构等近 40 个数据库. 作为 EMBL 的节点, 它建立了镜相系统, 具有多种服务功能.

在各综合数据库中, EST 的增长速度最快. 1991 年各个公共数据库中的 EST 数目不足2 000 条, 到 1999 年 12 月底, 在最大的数据库 GenBank 中的 EST 数量已增至3 436 681条 (http://www.ncbi.nlm.nih.gov/dbEST/dbEST-summary.html). 数据库的发展除了具有增长与更新速度快、复杂度与网络化程度高的特点外, 数据库的专门化也是一趋势, 即建立特定物种或特定研究内容的数据库[17]. 如作物研究数据库 GrainGenes, 包括小麦、玉米、大豆基因组子数据库等. 拟南芥和水稻由于其在基因组研究中的特殊作用, 也有许多独立的数据库或子数据库.

3.3 EST 序列分析

在 EST 研究中, 使用最多的方法就是序列相似性比较, 以此来确定 EST 的功能. BLAST (Basic Local Alignment Search Tool) 是应用较广的工具软件之一, 为同源分析的软件包, 包括有 BLASTN, BLASTP, TBLASTN, TBLASTX, BLASTX 等 5 个软件[18]. 其中 BLASTN、BLASTX 和 TBLASTX 是为核酸序列查询设计的, 在 EST 分析中都要用到; BLASTP 和 TBLASTN 是为蛋白质序列查询设计的. BLASTN 是待查询序列及其互补序列一起对库查询, 速度快但灵敏度不高. BLASTX 是将待查询序列按 6 种可能的阅读框架进行翻译, 然后对蛋白质序列库进行查询, 该软件灵敏度高、对序列错误的忍耐性也大. 在进行 EST 分析时, 同时使用 BLASTN 和 BLASTX 会得到较准确的结果. TBLASTX 是对待查序列和所有核酸序列库中的序列进行 6 个读码框的翻译, 然后在蛋白质水平上比较. 由于该软件运用计算资源量大, 检索比较昂贵, 所以用其查询只限于对 dbEST.

4 植物 EST 计划中常遇的几个问题

4.1 丰度问题

用构建的普通 cDNA 文库进行测序, 其中高丰度的基因将会被反复测序, 所以丰度问题是各 EST 计划中均会遇到的一个棘手问题. 如水稻 EST 计划中总共测定 27 428 个克隆, 重复的为 10 507个, 平均重复度为 57.5%[19]; 拟南芥 EST 计划所测得 1 152 个克隆中, 31% 为重复的序列[20]. 对于为寻找新基因或研究基因差异表达而言, 用这样的 cDNA 文库进行 EST 测序, 一方面稀有基因容易遗漏, 再者重复测序也造成人、财、物的巨大浪费. 解决这一问题一般有三种措施: a. 建立均一化的 cDNA 文库, 使 cDNA 文库中基因出现的频率基本一致. b. 建立差减文库, 扣除管家基因编码的产物, 使要研究的差异基因富集. c. 将出现频率高的基因标记探针, 文库经杂交筛选后再测序.

4.2 测序方向

cDNA 的 5′ 与 3′ 端均可作为测序起点, 选择哪端作为测序起点, 则有几个问题需要考虑: 一是 EST 编码蛋白质的信息应满足同源序列比较分析; 二是决定于用 EST 来进行研究的目的. 5′ 端非编码区小, 所含信息多, 一般在寻找新基因或研究基因差异表达时用 5′ 端 EST 较好, 大部分 EST 计划都是选用 5′ 端进行测序的. 3′ 端 mRNA 有一 20~200 bp 的 polyA 结构, 同时靠近 polyA 又有特异性的非编码区, 所以从 3′ 端测得 EST 含有编码的信息较少. 但研究也表明, 10% 的 mRNA3′ 端有重复序列, 这可以作为 SSR 标记; 非编码区有品种的特异性, 可以作为 STS 标记[1,19].

4.3 关于新基因

EST 研究中有相当一部分为首次发现的基因, 这就是所谓新基因. 新基因在植物 EST 测定中比例很高, 这与植物基因组研究开展得晚, 已知功能基因相对少有关. 如拟南芥 1993 年所测的1 152个 EST 中 68% 的为新基因[20]; 1997 年约 1.5 万个不

重复 EST 中 60% 为未知的[21]. 水稻 EST 研究, 1996 年所测 EST 中约有 75% 为新基因[19]; 1998 年未知部分则仍为 75%[20]. 由于 EST 中存在大量未知序列, 因此用 EST 技术进行基因差异表达研究, 特别是有关植物生理生化机理研究时, 常常遇到困难. 相信随着植物基因鉴定工作的不断进行, 利用 EST 技术分析基因功能将会成为未来鉴定基因功能的主要方式.

参 考 文 献

1 Hatey F, Tosser-Klopp G, Clouscard-martinato C, *et al*. Expressed sequenced tags for genes: a review. Genet Sel Evol, 1998, **30** (5): 521~ 541

2 Adams M D, Kelley J M, Gocayne J D, *et al*. Complementary DNA sequencing: expressed sequence tags and human genome project. Science, 1991, **252** (5013): 1651~ 1656

3 Rounsley S, Linx K K. Large scale sequencing of plant genome. Curr Opin Plant Biol, 1998, **1** (2): 136~ 141

4 Sasaki T. The rice genome project in Japan. Proc Natl Acad Sci USA, 1998, **95** (5): 2027~ 2028

5 Hisada S, Akinama T, Edo T, *et al*. Expressed sequence tags of citrus fruit during rapid cell development phase. Amer Soc Hort Sci, 1997, **122** (6): 808~ 812

6 Desprez T, Amselem J, Caboche M, *et al*. The Arabidopsis thaliana cDNA sequencing project. Plant J, 1998, **14** (5): 643 ~ 652

7 Allona I, Quinn M, Shoop E, *et al*. Analysis of xylem formation in pine by cDNA sequencing. Proc Natl Acad Sci USA, 1998, **95** (16): 9693~ 9689

8 Gilpin B J, Mccallum T A, Frew T J. A linkage map of the pea (pisum stativum L) genome containing cloned sequences of known function and expressed sequence tags (ESTs). Theor Appl Genet, 1997, **95** (8): 1289~ 1299

9 Harushima Y, Yano M, Shomura A, *et al*. A high-density rice genetic linkage map with 2275 markers using a single F_2 population. Genetic, 1998, **148** (1): 1~ 16

10 Liu J Y, Hara C, Umeda M, *et al*. Analysis of randomly isolated cDNAs from developing endosperm of rice (*Oryza sativeo* L): evalution of expressed sequence tags and expression levels of mRNAs. Plant Molecular Biology, 1995, **29** (4): 685~ 689

11 Epple P, Apel K, Bohlmann H. ESTs reveal multigene family for plant dfensins in *Arabidopsis thalina*. FEBS Letters, 1997, **400** (2): 168~ 192

12 Sterky F, Regan S, Karlsson J, *et al*. Gene discovery in the wood-forming tissue of poplar: Annalysis of 5692 expressed sequence tags. Proc Natl Acad Sci USA, 1998, **95** (22): 13330~ 1335

13 Green P, Lipman D, Hittier L, *et al*. Ancient conserved region in new gene sequence and protein database. Science, 1996, **259** (5102): 1711~ 1716

14 Benton D. Bioinformatics-principles and potential of a new multidisciplinary tool. TIBTECH, 1996, **14** (8): 261~ 272

15 Leipe D D. Genome and DNA sequence database. Curr Opin GenDevel, 1996, **6** (6): 686~ 691

16 罗静初, 江 涛, 李 兵, 等. 分子生物信息镜象系统和数据库. 高技术通讯, 1998, **10** (10): 61~ 63, 53
Luo J C, Jiang T, Li B, *et al*. Mirror system of molecular bio-information and database. Advanced Technology Communication, 1998, **10** (10): 61~ 63, 53

17 Gelbart W M. Database in genetic research. Science, 1998, **282** (5389): 659~ 661

18 夏 云, 雷二庆, 王槐春. Internet 实用技术与生物医学应用. 北京: 军事医学出版社, 1997. 341~ 354
Xia Y, Lei E Q, Wang H C. Internet technique and it's application in bio-medcine. Beijing: Millitary Medcine Publishing House, 1997. 341~ 354

19 Yammanoto K, Sasaki T. Large-scale EST sequencing in rice. Plant Molecular Biology, 1997, **35** (1): 135~ 144

20 Hoog C. Isolation of a large number of novel mammalian genes by a differential cDNA library screening strategy. Nucleic Acids Research, 1991, **19** (22): 6123~ 61127

21 Delseny M, Coole K, Raynaol M, *et al*. The Arabidopsis thaliana cDNA sequencing project. FEBS Letter, 1997, **403** (3): 221~ 224

Progress in Expressed Sequence Tags (EST) Project of Plant Genome*

LUO Meng**, JIA Ji-Zeng

(Key Laboratory of Crop Germplasm & Biotechnology/ Chinese Agriculture Ministry, Institute of Crop Germplasm Resources, Chinese Academy of Agricultural Sciences, Beijing 100081, China)

Abstract Plant expressed sequence tags (EST) project is a new research area in plant genome. By a single pass large scale sequencing of cDNA libraries, ESTs can be acquired and used to analyze gene expression, organization, construction. Some kinds of plant genome projects, the major research in plant EST project, the fuction of bioinformatics in EST analysis, dbEST and inquiry service, some problems in EST research are discussed.

Key words plant genome, expressed sequence tags, bioinformatics

* This work was supported by grants from 973 (973-08-02) Project and National Nature Science Foundation of China (399800029).

** Corresponding author. Tel: 86-10-62186652, E-mail: lmzy@public2.east.net.cn

Received: September 27, 2000 Accepted: November 3, 2000