林业科学研究 2004,17(6):804～809
Forest Research

文章编号:1001-1498(2004)06-0804-06

表达序列标签(EST)分析及其在林木研究中的应用

李 虹[1,2], 卢孟柱[2], 蒋湘宁[1]
(1. 北京林业大学,北京 100083; 2. 中国林业科学研究院林业研究所,北京 100091)

摘要:简要叙述了表达序列标签 EST 技术的原理和流程,综述了 EST 在研究林木木材形成和其它生物学过程时新基因的发现、基因表达分析和基因芯片方面的应用进展以及在开发林木单核苷酸多态性和简单序列重复等分子标记和构建遗传图谱方面的应用进展,并对其在林木基因组研究中的应用前景进行了展望。

关键词:EST;新基因发现;基因表达;分子标记
中图分类号:Q78 **文献标识码**:A

1991 年 Adams 等人从三种人脑组织的 cDNA 文库中随机挑取 609 个克隆进行测序,从而得到一组人脑组织的表达序列标签 EST (expressed sequence tags),并将其与数据库进行序列同源性对比,结果表明:该组 EST 中有 36 个代表已知基因,337 个代表未知基因,这是关于 EST 技术应用的首次报道,并首次提出了 EST 的概念[1]。随着人类基因组计划的顺利进行,EST 技术首先被广泛应用于寻找人类新基因,绘制人类基因组图谱,识别基因组序列编码区等研究领域,之后又被广泛应用于植物基因组研究[2]。随着 EST 测序的飞速发展,到 2003 年 6 月,美国国家生物技术信息中心(NCBI)的 EST 数据库中(dbEST) (http://www.ncbi.nlm.nih.gov/dbEST/index.html)已录入的来自不同物种的不同组织的 EST 共有 17 291 123 条,其中人和鼠的最多。EST 也被广泛应用于新基因的发现、基因鉴定、基因克隆、构建基因组图谱、基因定位分析、基因表达分析等方面。在植物方面,除了拟南芥(*Arabidopsis thaliana* (L.) Heynh.)、水稻(*Oryza sativa* L.)、小麦(*Triticum aestivum* L.)、大麦(*Hordeum vulgare* L.)、大豆(*Glycine max* (L.) Merr.)、玉米(*Zea mays* L.)、棉花(*Gossypium herbaceum* L.)等模式植物和农作物以外,近年来也开展了一些木本植物的 EST 研究,首先报道的是火炬松(*Pinus taeda* L.) EST 分析,随后是杂交杨(*Populus tremula* L. ×*P. tremuloides* Michx.)和毛果杨(*P. trichocarpa* 'Trichobel.')等其它林木。

1 EST 技术的原理和步骤

EST 指从不同组织来源的 cDNA 文库中随机挑选克隆,进行 5′或 3′端测序后得到的部分 cDNA 序列,一个 EST 对应于某一种 mRNA 的 cDNA 克隆的一段序列,长度一般为 300～500

收稿日期: 2003-08-06
基金项目: 国家重点基础研究规划项目(973)"树木育种的分子基础研究"(G1999016000)
作者简介: 李虹(1974—),女,湖南益阳人,硕士.

bp[1]。cDNA 是由来源于某一组织的 mRNA 在体外经逆转录酶逆转录合成单链,再由 DNA 聚合酶等催化合成双链,只含有基因编码区域,因此,EST 是了解基因表达的"窗口",可代表生物体某种组织某一时间的一个表达基因,故被称之为"表达序列标鉴";而且 EST 的数目可以显示所代表的基因表达的拷贝数,一个基因的表达次数越多,其相应 cDNA 克隆也就越多,所以通过对 cDNA 克隆的测序分析可以了解基因的表达丰度。EST 技术的具体流程为:(1)从组织细胞中提取 mRNA,构建标准 cDNA 文库;(2)从中获得大量的单个 cDNA 克隆;(3)碱裂解法或 PCR 扩增制备测序模板;(4)cDNA 片段 5 端或 3 端 300～500 碱基的测序;(5)将测序所得的 EST 序列与 dbEST 等数据库中已知的核酸和蛋白质序列进行同源性比较分析,可以鉴定出哪些代表已知基因,哪些代表未知序列,后者可能代表新基因,并进行基因表达丰度分析,确定这些基因在该组织中的表达水平;(6)新基因及未知基因的基因库登录。目前 cDNA 文库构建都有现成的试剂盒,方法成熟,同时 DNA 测序技术的飞速发展,进一步降低了大规模 DNA 序列测定的成本。EST 数据库构建费用的成倍降低为林木基因组学研究的开展提供了良机。

2　EST 在林木研究中的应用

近年来,随着一些木本植物 EST 分析工作的启动,使数据库中木本植物 EST 的数目越来越多,这些 EST 为新基因的发现和基因表达研究提供了大量的信息和分析材料,也为高密度林木遗传图谱的构建所需分子标记的开发奠定了基础。

2.1　新基因的发现

利用 EST 技术分析得到的基因主要有三种:第一是已知基因,是为人类已鉴定和了解的基因;第二是以前发现但功能未经鉴定的基因,但根据组织发育特点可以推测该基因的功能;第三是未知基因,即该基因在数据库中无同种或异种基因的匹配;所以利用 EST 技术不但可迅速地确定部分基因的功能,而且为推测未知功能基因和发现新的基因提供了重要基础。EST 分析是基于大量基因测序基础上,具有基因组学的研究特点,为像树木这种分子生物学研究背景少、突变体难以获得的植物提供了有效研究手段。1998 年 Allona 等[3]构建了火炬松未成熟木质部的 cDNA 文库并从中获得了 1 097 个 EST 序列,通过与公共数据库序列同源对比发现 59 %与已知功能的基因序列相似,其中大约 10 %为编码细胞壁形成有关的因子,如一些参与细胞壁形成的蛋白质、已知的木质素生物合成的酶类和几个与糖类代谢相关的酶,另外还有许多是推测的调节蛋白。2002 年 1 月在圣地亚哥召开的动植物和微生物基因组学会议上,来自美国北卡罗来纳州立大学的 Johnson 等[4]阐述了美国 1999 年启动了从基因组学途径研究火炬松木材形成的分子基础的计划,目前已构建了 4 种不同的正在发育的木质部区域 cDNA 文库并建立了 EST 库,总共 60 000 多个序列,约 80 %产生了有效的 ESTs,大量的火炬松 EST 与已知的植物基因有高度的同源性,其中相当一部分与细胞壁形成有关。在 EST 单基因克隆库里有一部分是新基因,表现出与拟南芥或其它植物的序列没有明显的同源性。1998 年瑞典的 Sterky 等[5]从杂交杨的形成层区域和毛果杨的未成熟木质部区域获得了 5 692 个 ESTs 序列,通过与公共数据库同源对比,发现形成层 EST 库的 63 %和木质部 EST 库的 54 %与 820 种已知功能蛋白质序列相似,两个文库中分别有 25 %和 37 %的 ESTs 与来自其他物种的且功能未知的序列有显著同源性。另外,12 %和 9 %的 ESTs 与公共数据库中的任何序列均无相似性,表明这些序列代表的基因可能为新基因,并在木材形成中具有特殊功能。上述研究是杨树(*Populus*

spp.)基因组计划的一部分,到现在杨树 EST 已增加到 95 000 多个,分别来自不同组织和发育阶段的 20 个 cDNA 文库。分析表明这些 EST 来自杨树基因组可能编码基因总数 40 000～50 000中的 15 000～20 000 个基因。所有这些 EST 的功能归类还没全部完成,但其中几个子集已被分析。例如杨树幼嫩叶片 36 %的 ESTs 为与能量代谢有关的基因,而衰老叶片中与细胞程序化死亡和蛋白质降解相关的 ESTs 占的比例比幼嫩叶片增加 2～3 倍[6],因此,EST 分析能够检测基因表达的趋势和揭示特定组织的生物学过程。此外,Hisada 等[7]对温州蜜桔(*Citrus unshiu* Marc.)果实细胞快速膨大期的幼果组织和上岛脐橙(*Citrus sinensis* (L.) Osbeck)的未成熟种子[8]进行了 EST 分析;在动植物和微生物基因组学会议上还报道对洋槐(*Robinia pseudoacacia* L.)[9]、海岸松(*P. pinaster* Ait.)[10]、桉树(*Eucalyptus globulus* Labill.)[11]等也进行了类似的 EST 研究。

2.2 基因表达研究

EST 技术稳定性高,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,从而能在整体水平研究相关的功能和代谢。如 Sterky 等[5]通过比较杨树形成层区域和发育木质部的 EST 库发现:两文库包含不同的高丰度的转录产物,木质部库中高丰度转录产物的比率高于形成层库;木质部库中细胞壁相关基因的表达几乎是形成层库的 2 倍,而蛋白质合成相关基因的表达是形成层库的一半;两库中都有木质素生物合成基因,但在木质部库中丰度更高。与木质化有关的其它基因的表达也有明显区别,特别是漆酶、S-腺苷蛋氨酸合成酶和过氧化物酶的丰度,过氧化物酶在形成层库中表达更高,而漆酶、S-腺苷蛋氨酸合成酶在木质部库中高度表达。S-腺苷蛋氨酸合成酶被认为是通用甲基供体,在木质素单体合成中非常重要。虽然漆酶、过氧化物酶都参与木质素单体聚合反应,但漆酶的作用更重要,其在火炬松的高水平表达也证明了这一点。

EST 除了通过丰度分析可以确定基因表达水平外,还可用于制备 DNA 芯片,利用不同组织和发育时期的试验材料进行基因表达研究,成为鉴定新基因和功能的初始材料。利用 EST 序列,采用 PCR 技术可以方便地扩增代表不同基因的 cDNA 片段,用于制备基因芯片。Johnson 等[4]在 2002 年召开的动植物和微生物基因组学会议上叙述了用一部分火炬松 EST 制备芯片,研究了幼材和成材在正常生长和受到机械压迫时以及早材和晚材形成过程中的基因表达变化。Hertzberg 等[12]利用来自杂交杨的 2 995 个 EST 制备芯片,研究了木材形成过程中的几个阶段如细胞分化、扩张、次生壁形成、木质化和细胞程序性死亡的基因表达变化,揭示了一些编码木质素和纤维素生物合成的基因、木质化过程的许多转录因子和其它潜在的调节因子受严格的特定发育阶段的转录调节。目前在上述杨树 95 000 个 EST 的基础上,开始了高密度芯片的制备,如瑞典进行的杨树基因组计划中制备了一个包含 13 000 个 EST 的芯片,这些 EST 来自于 35 000 个 cDNA 克隆测序后的单基因集,用于研究杨树的许多生长发育过程中基因的表达、鉴定和功能分析[6]。作为合作者之一,Taylor 等[6]利用该高密度杨树 EST 的芯片研究了长期处于高浓度 CO_2 中杨树基因的表达,发现其中 1 500 个 EST 表达增加,而另外 1 000 个 EST 则表现出表达下降,Taylor 强调这些研究将揭示在未来气候变化下树木对环境适应的机制。

2.3 开发分子标记和图谱的构建

EST 片段由于其多态性高,可以开发为分子标记,用于林木群体的遗传分析,大量的 EST 分子标记可以用于建立遗传连锁图谱。EST 序列本身的碱基变化就可以开发单核苷酸多态性

标记(SNP),可以采用 PCR 扩增结合测序或梯度变性电泳加以鉴定。微卫星 DNA 是由少数几个核苷酸(一般为 2～4 个)为单位多次串联重复的 DNA 序列,故又称为简单序列重复(Simple sequence repeats (SSR)),主要是以两个核苷酸对为重复单位。SSR 在基因组中非常丰富,所以 EST 中也存在 SSR,设计 SSR 两侧保守区引物,通过 PCR 扩增就能检测出 EST 中的 SSR,所以利用 EST 可以开发出 SSR 标记。Xu Yong 等[13]利用扁桃(*Amygdalus communis* L.)种质的 1 057 个 EST 设计了 26 个 SSR 引物,通过扩增获得了 11 个 SSR 标记,这些 SSR 标记被用于研究扁桃种质的遗传多样性,并发现来源于基因组的 SSR 和来源于 EST 的 SSR 在每个位点上检测到的等位基因的数目有很大区别。Scott 等[14]分析 5 000 个葡萄(*Vitis* spp.) EST 得到 124 个微卫星 DNA,从中设计 16 个 SSR 引物,通过扩增获得 10 个 SSR 标记,检测了它们的多态性和可转移性,并与来源于基因组的 SSR 进行了比较。Decroocq 等[15]从杏(*Prunus armeniaca* L.)和葡萄的 EST 中得到了一些 SSR,并研究了来源于 EST 的 SSR 在葡萄科(Vitaceae) 和蔷薇科(Rosaceae)之间转移的可能性。

利用 EST 开发的分子标记,有如下优点:(1)如果一个 EST 标记被发现与一个有意义的遗传性状有关,那么这个 EST 所代表的基因就有可能直接影响这个性状;(2)与候选基因同源或在某个组织中有差异表达的 EST 可被选定为遗传作图的标记,对了解分析目标性状大有益处;(3)由于 EST 来源于编码区 DNA,一般有高度的序列保守性。与多数来自非表达区的其它标记如 AFLP、RAPD、SSR 相比,EST 标记更可能在家族和物种之间转换,因此,EST 标记在远亲物种之间校正基因组连锁图谱和比较数量性状定位方面特别有用;同样,如果一个目标物种缺少 DNA 序列信息,那么其它物种的 EST 能被用作这个物种的遗传作图,因此利用 EST 遗传作图将使物种之间连锁信息的转换更快,能用作校正标记,实现多个图谱整合,并进行比较基因组学研究。Temesgen 等[16]利用构建的火炬松幼苗针叶和幼树木质部两个 cDNA 文库,得到了部分 EST,其中有 56 个 EST 标记可定位到由两个作图群体构建的火炬松遗传连锁图上和一张整合的火炬松遗传连锁图上。不像常用的其它分子标记,EST 可以定位已知功能的基因或定位影响火炬松重要性状的候选基因。Brown 等[17]利用火炬松的 90 个 EST 开发的标记,对松属(*Pinus* L.)的单维管束亚属(*Strobus*)与双维管束亚属(*Pinus*)和松科的花旗松(*Pseudotsuga menziesii* (Mirbel.) Franco)进行了比较作图研究,结果表明:89 %、49 %和 22 %的 EST 引物可以从双维管束亚属、单维管束亚属及花旗松中实现扩增,多态性比例处于 37 %～61 %。35 个 EST 标记处于火炬松和花旗松图谱的相同位点,为构建松属的"通用"图谱、研究基因组的结构与进化奠定了基础。Komulainen 等[18]利用 EST 标记和其它分子标记构建了欧洲赤松(*P. sylvestris* L.)的遗传图谱,这些 EST 标记的一部分以前已被用于火炬松的作图,并比较了欧洲赤松和火炬松的基于 EST 的遗传图谱。此外,对辐射松(*P. radiata* D. Don)[19]、桉树[20]和云杉(*Picea asperata* Mast.)[21]等也利用 EST 进行了遗传作图。遗传图谱的构建是基因组研究中的重要环节,可为基因定位及基因组结构和功能的研究打下基础,高密度的遗传图谱有助于克隆基因和精确地解析数量性状基因。

3　展望

EST 技术在林木基因组研究中的应用展示了良好的前景。虽然林木基因组研究开展得较晚,树种也有限,与人、拟南芥和水稻等作物的基因组研究比较起来相对滞后,但随着 EST 分

析技术的日趋完善、其应用的不断开发,相信其在林木研究中会越来越被重视,必将在林木基因组学中发挥重要作用。另外,分子育种对品质、抗性基因的定位和基因的分离需求紧迫,EST 分析无疑可以用来研究这些性状形成过程的基因及其表达,同时为基因的表达分析提供基础(基因芯片),成为阐明目标性状分子机理的重要技术途径。

EST 序列代表着染色体上识别位点,它对应的序列标记位点(sequence-tagged site, STS)已成为绘制基因组物理图谱的标准标记,可以突破传统林木遗传图谱密度低、通用性差等应用瓶颈,所以 EST 标记不仅用于"通用"林木遗传图谱的构建,也将为林木基因组物理图谱的构建奠定基础,物理图谱的构建也为基因组结构分析、克隆已经定位的基因提供了条件。随着模式树种杨树基因组测序工作的完成和功能基因组计划的实施,预见未来几年 EST 在基因识别、基因表达和功能研究等方面将发挥越来越大的作用,而且随着生物信息学的发展,在林木研究中的应用范围也将更加广阔。

我国有许多特有木本植物资源急待开发,EST 分析作为基因组研究的首要手段,无疑是新基因发掘的良好工具,这应引起我国科学工作者的高度重视。

参考文献:

[1] Adams M D, Kelly J M, Gocayne J D, et al. Complementary DNA sequencing: expressed sequence tags and human genome project[J]. Science, 1991, 252: 1651～1656

[2] 骆蒙, 贾继增. 国际麦类基因组 EST 计划研究进展[J]. 中国农业科学, 2000, 33(6): 110～112

[3] Allona I, Quinn M, Shoop E, et al. Analysis of xylem formation in pine by cDNA sequencing[J]. Proc Natl Acad Sci USA, 1998, 95: 9693～9698

[4] Johnson A, Kinlaw C, Loopstra C, et al. A genomic approach to wood formation in loblolly pine[R]. Plant, Animal & Microbe Genomes X Conference, San Diego, 2002

[5] Sterky F, Regan S, Karlsson J, et al. Gene discovery in the wood forming tissues of poplar: Analysis of 5 692 expressed sequence tags[J]. Proc Natl Acad Sci USA, 1998, 95: 13330～13335

[6] Wullschleger S, Jansson S, Taylor G. Genomics and Forest Biology: Populus emerges as the perennial favorite[J]. Plant Cell, 2002, 14: 2651～2655

[7] Hisada S, Akihama T, Endo T, et al. Expressed sequence tags of citrus fruit during rapid cell development phase[J]. J Amer Soc Hort Sci, 1997, 122(6): 808～812

[8] Hisada S, Moriguchi T, Hidaka T, et al. Random sequencing of Sweet Orange (*Citrus sinensis* Osheck) cDNA library derived from young seeds[J]. J Japan Hort Sci, 1996, 65(3): 487～495

[9] Han K H, Yang J, Park S, et al. Genomics of wood formation in Black Locust[R]. Plant & Animal Genomes IX Conference, San Diego, 2001

[10] Frigerio J M, Dubos C, Plomion C, et al. Gene expression in shoots and roots of well watered and drought-stressed Maritime Pine seedlings[J]. Plant & Animal Genomes Ⅷ Conference, San Diego, 2000

[11] Sawbridge T I, Drenth J, Hallinan, et al. EST sequencing in Eucalyptus[R]. Plant & Animal Genome Ⅶ Conference, San Diego, 1999

[12] Hertzberg M, Aspeborg H, Schrader J A. Transcriptional roadmap to wood formation[J]. Proc Natl Acad Sci USA, 2001, 98: 14732～14737

[13] Xu Y, Ma R C, Cao M Q, et al. Genetic diversity and phylogenetic analysis of Almond Germplasm using EST-SSRs and genomic SSRs [R]. Plant & Animal Genomes XI Conference, San Diego, 2000

[14] Scott K D, Eggler P, Seaton G, et al. Analysis of SSRs derived from grape ESTs[J]. Theor Appl Genet, 2000, 100: 723～726

[15] Decroocq V, Fav M G, Hagen L, et al. Development and transferability of apricot and grape EST microsatellite markers across taxa[J]. Theor Appl Genet, 2003, 106: 912～922

[16] Temesgen B ,Brown G R ,Harry D E ,et al. Genetic mapping of expressed sequence tag polymorphism (ESTP) markers in loblolly pine (*Pinus taeda* L.) [J]. Theor Appl Genet ,2001 ,102 :664～675

[17] Brown G R ,Kadel Ⅲ E E ,Bassoni D L ,et al. Anchored Reference Loci in Loblolly Pine (*Pinus taeda* L.) for Integrating Pine Genomics[J]. Genetics ,2001 ,159 :799～809

[18] Komulainen P ,Brown G R ,Mikkonen M ,et al. Comparing EST-based genetic maps between *Pinus sylvestris* and *Pinus taeda*[J]. Theor Appl Genet ,2003 ,107 :667～678

[19] Cato S A ,Gardner R C ,Kent J ,et al. A rapid PCR-based method for genetically mapping ESTs[J]. Theor Appl Genet ,2001 ,102 :296～306

[20] Thamarus K A ,Groom K,Murrell J ,et al. A genetic linkage map for *Eucalyptus globules* with candidate loci for wood ,fibre ,and floral traits[J]. Theor Appl Genet ,2002 ,104 :379～387

[21] Gosselin I ,Zhou Y ,Bousquet J ,et al. Megagametophyte-derived linkage maps of white spruce (*Picea glauca*) based on RAPD ,SCAR and ESTP markers[J]. Theor Appl Genet ,2002 ,104 :987～997

Expressed Sequence Tag (EST) and Its Application in Forest Research

LI Hong[1,2] , *LU Meng-zhu*[2] , *JIANG Xiang-ning*[1]

(1. Beijing Forestry University ,Beijing　100083 ,China ; 2. Research Institute of Forestry ,CAF ,Beijing　100091 ,China)

Abstract: This article has introduced the basic principles and procedure of EST analysis ,and also reviewed the application of EST in novel gene finding ,gene expression analysis and the use in preparation of gene-chips in genomic studies on different biological processes ,wood formation for instance. EST database provide a resource to develop molecular markers such as SNP ,SSR ,etc. ,the latter has been used to construct genetic linkage maps. It's prospect of its application on forest genomics is also discussed.

Key words: EST; gene identification ; gene expression analysis ; molecular marker