

利用生物信息数据库克隆玉米基因 *PIS2* 与 *MFSP* 全长^①

岳桂东^② 隋镇华 李 煦 张举仁^③

(山东大学生命科学院 济南 250100)

摘 要 利用玉米表达序列标签数据库(dbEST)电子克隆了玉米基因 *PIS2* 的 cDNA 全长,并采用反转录聚合酶链反应(RT-PCR)技术得到了包含该基因完整开放阅读框(ORF)的 cDNA 序列;借助水稻非冗余核酸数据库联合应用电子克隆技术和 RT-PCR 技术克隆得到玉米主易化子超家族蛋白基因 *MFSP* 的 cDNA 全长,并进一步利用植物基因组数据库(PlantGDB)快速地获得了该基因的基因组全长序列。研究证实,充分利用分子生物信息数据库,可以简便地克隆得到玉米目的基因的 cDNA 乃至基因组全长序列。

关键词 电子克隆,反转录聚合酶链反应(RT-PCR),玉米

0 引 言

电子克隆^[1](*in silico* cloning)是指利用计算机技术,依托现有的分子生物信息数据库表达序列标签(expressed sequence tag, EST)数据库、核苷酸数据库、基因组数据库等)网络资源,采用生物信息学方法(包括同源性检索、聚类、序列拼接等),通过 EST 或基因组序列的组装和拼接,利用反转录聚合酶链反应(RT-PCR)快速地获得部分乃至全长基因序列的一种方法^[2-4]。与传统的基因克隆方法相比,基于 EST 的电子克隆技术具有速度快、投入低、针对性强等优点,是克隆新基因的便捷方法之一^[5-7]。近几年来,随着模式生物基因组测序及生物信息学的快速发展,产生了大量可供利用的基因组信息和分析软件,充分利用这些数据和工具来寻找新基因,了解基因的功能,是功能基因组学研究的主要任务,也是目前基因克隆的主要策略之一^[8]。

在高等植物中,拟南芥和水稻的全基因组序列测序工作已经完成^[9-11],其他物种如麦类基因组计划、玉米基因组计划也正在进行,越来越多的分子生物信息数据库也相应得以建立或充实,我们可以充分利用这些生物信息数据库来辅助克隆我们感兴趣的基因序列,然而利用电子克隆技术克隆玉米等大基因组物种基因的工作报道很少^[12]。本文报道了分别利用玉米 EST 数据库克隆玉米磷脂酰肌醇合

成酶基因 *PIS2* 的 cDNA 全长,借用水稻核酸数据库克隆玉米主易化子超家族蛋白(major facilitator superfamily protein, MFSP)长基因 *MFSP* 的 cDNA 全长,以及利用植物基因组数据库获得玉米基因 *MFSP* 的基因组全长序列的研究。

1 材料与方法

1.1 利用玉米 EST 数据库克隆玉米基因 *PIS2* 的 cDNA 全长

1.1.1 数据库与软件

核酸数据库为美国国立生物信息中心(National Center for Biotechnology Information,简称 NCBI)的 GenBank 数据库(<http://www.ncbi.nlm.nih.gov/Database/index.html>);同源性分析程序使用 NCBI 网站(<http://www.ncbi.nlm.nih.gov>)的在线软件 BLASTN、BLASTX;ESTs 序列组装使用 DNASTAR 软件的 SeqMan 程序;开放阅读框(ORF)的寻找使用 NCBI 网上的 ORF Finder 程序。

1.1.2 电子克隆技术路线

A. 以一条 EST 序列为种子在 NCBI 上用 BLASTN 搜索玉米 EST 数据库,下载匹配序列。

B. 将下载的 ESTs 序列用 DNASTAR 软件进行第一次拼接和校对,得到第一个重叠群(contig)。

C. 以此重叠群为目标序列在 NCBI 数据库上重新搜索匹配序列,并下载匹配序列。

① 863 计划(2006AA10A107)资助项目。

② 男,1979 年生,博士;研究方向:玉米基因组学。

③ 通讯作者, E-mail: jrzhang@sdu.edu.cn
(收稿日期:2007-12-23)

D. 再将匹配序列进行拼接。

E. 重复以上步骤,直到不能再延伸为止。

F. 将最后得到的 contig 作为目标序列在 NCBI 上用 ORF Finder 程序寻找 ORF。

1.1.1.3 植物材料、菌株和质粒

植物材料为玉米 (*Zea mays* L.) 优良自交系 DH4866; T/A 克隆载体 pGEM® T-Easy Vector Systems 购自 PROMEGA 公司; DNA Marker DL2000 和胶回收试剂盒购自 TAKARA 公司; *E. coli* DH5 α 由本实验室保存。

1.1.1.4 玉米 cDNA 的合成

玉米材料培养至三叶期,使用 Trizol 试剂(上海生工公司)从叶片组织中提取总 RNA,利用 Oligotex® mRNA Kit (QIAGEN 公司)分离 mRNA 和 Smart™ RACE cDNA Amplification Kit (CLONTECH 公司)反转录合成 cDNA。

1.1.1.5 *PIS2* 基因的电子克隆

本实验室已经成功克隆了玉米磷脂酰肌醇合成酶(phosphatidylinositol synthase)基因 *PIS* 的 cDNA 全长(GeneBank 登录号为 AY370763),Southern blot 杂交分析发现该基因在玉米基因组中存在两个拷贝^[13]。以 *PIS*(AY370763)为种子序列检索玉米 EST 数据库,将得到的匹配序列进行拼接,得到了两条 contig 序列,其中一条 contig 是 AY370763,另一条 contig 长为 1346 bp,与 AY370763 的相似性为 94%,应是玉米基因组中第二个 *PIS* 基因,故命名为 *PIS2*。

1.1.1.6 RT-PCR 克隆 *PIS2* 基因

根据 1.1.5 中电子克隆的玉米基因 *PIS2* 的核酸序列设计一对引物(上游引物: 5'-GCCAGCAA-CAAGTTCACCTA-3'; 下游引物: 5'-ACAAAT-CAATAGCCGCCCTTC-3'),以 1.1.4 中合成的玉米 cDNA 为模板使用 PrimeSTAR 高保真 DNA 聚合酶(TAKARA 公司)进行 PCR 扩增。PCR 产物经胶回收试剂盒(TAKARA 公司)回收后连接 pGEM® T-Easy 载体,转化 *E. coli* DH5 α ,阳性重组子克隆送往上海博尚公司测序。

1.2 借用水稻核酸数据库克隆玉米基因 *MFSP* 的 cDNA 全长

1.2.1 *MFSP* 基因的电子克隆

本实验室已成功构建了 2 个干旱胁迫下开花期玉米叶片消减文库,并对这 2 个文库的部分克隆进

行了测序分析^[14]。有一条长 377bp 的 EST 在这 2 个文库中多次出现,对该 EST 进行 BLASTX 分析,发现其为玉米主易化子超家族蛋白(MFSP)基因 *MFSP* 的部分片段。

以 377bp 的 EST 为种子序列,经过电子克隆得到了一条长 883bp 的 contig。ORF Finder 软件分析发现该 contig 含有完整的 ORF 的 3'端,但其 5'端不完整。利用这条长 883 bp 序列在 NCBI 上检索水稻的非冗余核酸数据库(non-redundant database),找到了该序列在水稻中的同源基因序列(GeneBank 登录号为 AK119211)。进一步利用 AK119211 的 5'端序列去检索玉米的 EST 数据库,将 score 值最大的 4 条 EST(GeneBank 登录号分别为 DV516655, DT644466, DV031986, DR798562)组装成一条长 894bp 的 contig,分析发现其 ORF 的 5'端是完整的。

1.2.2 RT-PCR 克隆 *MFSP* 基因

根据 1.2.1 中两条 contig 序列设计特异性引物(上游引物 5'-GAGCCTCGCATCCCTAGT-3'; 下游引物 5'-GCAAGAAATGCTGAGGAGTAA-3'),以玉米 cDNA 为模板 PCR 扩增得到目的条带,经 T/A 克隆后送往上海博尚公司测序。

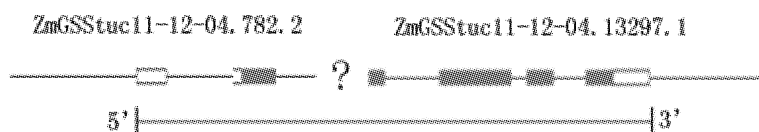
1.3 利用植物基因组数据库获得玉米基因 *MFSP* 的全长基因组序列

1.3.1 *MFSP* 基因组序列的电子克隆

登陆植物基因组数据库(<http://www.plantgdb.org/>),以 *MFSP* 基因的 cDNA 序列为种子,使用在线软件 BLASTN 检索该数据库中的玉米基因组数据,得到两段 DNA 序列: ZmGSStuc11-12-04. 782. 2 和 ZmGSStuc11-12-04. 13297. 1,经分析发现这两段序列分别是玉米基因 *MFSP* 基因组的上游和下游的两段序列。图 1 为 *MFSP* 的基因组序列。如图 1 所示, ZmGSStuc11-12-04. 782. 2 包含有 2 个完整的外显子和上游的启动序列;而 ZmGSStuc11-12-04. 13297. 1 则含有 4 个外显子和相应的内含子,但是它的最前面的外显子是不完整的。

1.3.2 未知序列的获得

根据 1.3.1 中两段 DNA 序列设计引物(上游引物 5'-TAACCGCAACTGCTTACCAA-3'; 下游引物 5'-AACCATCCAAGTGCCACG-3'),以 CTAB 法提取的玉米基因组 DNA 为模板 PCR 扩增得到目的条带, T/A 克隆后送往上海博尚公司测序。



黑框代表编码区域；空白框代表非编码区域；? 表示中间序列未知

图 1 *MFSP* 的基因组序列

2 结果

2.1 *PIS2* 基因全长 cDNA 序列的获得

图 2 为 *PIS2* RT-PCR 扩增电泳图。如图 2 所示,RT-PCR 得到了长 1.1kb 的主带,将该条带 T/A 克隆后测序分析发现其序列与 1.1.5 中电子克隆结果基本相符。玉米基因 *PIS2* 包含有一个长 648 bp 编码 215 个氨基酸的开放读码框,与已克隆的玉米 *PIS* 氨基酸水平相似性为 94%,现已将该基因提交至 GeneBank 数据库,获得的登录号为 EF103288。

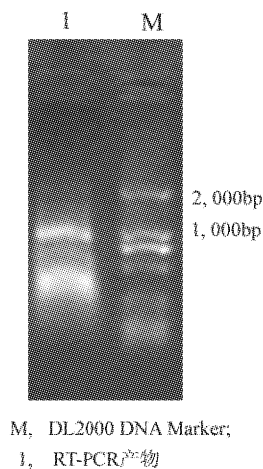


图 2 *PIS2* RT-PCR 扩增电泳图

2.2 *MFSP* 基因全长 cDNA 序列的获得

图 3 为 *MFSP* RT-PCR 扩增电泳图。如图 3 所示,利用 RT-PCR 的方法扩增出长 1.8kb 的目的条带,T/A 克隆后测序分析发现,该基因 cDNA 全长为 1995 bp(图 4),其中两条引物之间长 1804 bp 的序列已经 RT-PCR 测序验证。该 cDNA 包含一个长 1440 bp 的通读框,编码含 479 个氨基酸残基的蛋白质序

列。3'非翻译区长 273 bp,在 poly(A)上游的 26 bp 处有加尾信号 AATACA。将该 cDNA 序列在 GenBank 中使用 BLASTX 同源性检索对蛋白质序列进行比较分析,发现所得基因编码的蛋白与水稻的 *MFSP* 蛋白的氨基酸水平相似性很高(78%),由此推断该基因编码的蛋白是玉米中的一种 *MFSP* 蛋白,将该基因命名为 *MFSP* 并提交至 GeneBank 数据库,获得的登录号为 EF567316。

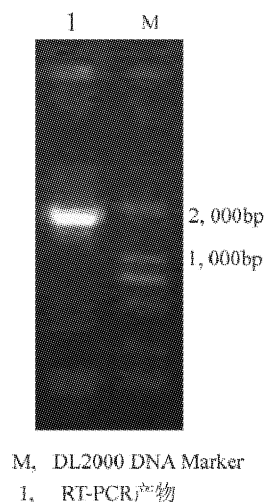


图 3 *MFSP* RT-PCR 扩增电泳图

2.3 *MFSP* 全长基因组序列的获得

将 1.3.2 中克隆测序所得的 DNA 序列与长为 2825bp 的 ZmGSStuc11-12-04.782.2 及长为 3566 bp 的 ZmGSStuc11-12-04.13297.1 进行电子拼接,得到长为 7112 bp 的 DNA 序列。该序列包含 *MFSP* 基因的完整基因组序列。经分析发现,*MFSP* 基因包含 6 个外显子,并且内含子和外显子交界区均符合 GT-AG 规则。该序列已提交至 GeneBank 数据库,获得的登录号为 EF608458。

5'-CAGGCGGCGGCTTGCCATTACCGTCCTCTCTTACCCGTCGTCCTTCACTGCAATTGAGCATTGCTTGTCTCCCTCCCGAC
 AAAAAACCCCTGTCCCGTCGAGCCTCGCATCCCTAGTTCCTTACAAGCAAGGAAGGAAACCCGGGCAACACGACCCGGCGGTGCCA
 GTGCCACGAAAATTAGCCGGCTCCCTCCCGACGTCGCGGGCGGCGAGCAACGGTGCTGGCTCCCAACATGGAGACATATACCACA
 GACGACGCGCTTACCATTATGGGGTTTGAAAATTCCAAGCGCTTGTTCTTTCTACGCTGGAATGGCTGGGTTCAGAAAGCTATGGAG
 CTCATGCTCTTGTCGTTTCTCGGGCCATTATACGAGAGGAGTGGAAACGTCCTCCCGAGAGAACGAGAGCCTCCTTTCGAGTGTGTCTTT
 GCTGGCATGCTGCTAGGTGCTGTGCTTGGGGCTTTGTTTCCGACAAATATGGACGGAGGACTGGTTTAATATTTCAACTCTATTACTAG
 TGAATGGGTTTTCTAAGCGCTTTATCTCTAACTACTTATGCTTGGTGGCTCTTCGGTTTCTGGTTGGAGTAGGAGTGGTGGCAGCCAT
 GTGTTTATATCTTGTTTTTGGAGTTTGTCTGCACAAAATCGTGGCACTTGGATGGTTATTTTCTCCCTTTCTGGACTCTCGGCACGATC
 TTGGAAGCTTCACTTGCATGGGTGCTATTGCCAGCATTGAATTGGAGGTGGCTGCTAGTCTTACCAGCCCTTCCATGCTTCTCTTGTCTCC
 TTTCTTTGGACTAACACCTGAGTCTCCGCGCTACCTTTGTGCACAAAATAGAATGCTGATGCCACATCTGTTCTAGAGAGAAATGGCCAAC
GCAAAACCAATCAGCTCTTCCCGCGGAGTTCTCAGTACAAATAGACAAAACAAATGTCACCAGCAGTTCTCCCTCCCGAGAGTGAATGI
CTTCTTCTGTCAGAGAGAAGGAATGCACAGTTGACAATGCCAAGAGCTCCGAATCTGGCAGTCTTGCTGCACTGGCAGTATTATCG
CGCGAAGTGTCTCAGATCAACTCTGCTTCTTTGGTTTGTCTTCTATGCAAAATCCTTTGCTTATTACGGGTAGTATTGCTGACATCCGAGTTG
 AGTGATGCAAAATAAAGTTGTGCATCTGGTGTGCTTTTGGTCTGCGCCAAAGGAAATCAATCTTTACAAAGACACTTTCATCACTAGTT
 TAGCAGAAATCCAGGTTTATTTTGTCTGCTGTTCTTGTGATTGGTTTGGTCCGAAAGCTTCGATGTGGTCTATGATGTTCCGATGCTGT
 GCTTTCCCTGGACCACTTGTACTTCAACAGACTGAGTTGTGACAACTATTTCTTATTTGGTGCCCGTGCATGTCCATGGGGAGCTTTAC
 AGTCTGTGCTCTTATGCTCCAGAGGTGTATCCAACATTGTCCGGTCAACCGGTGCAGGAATTGCGACCGCCATTGGTAGGATTGTTGGG
 GTTGTGTTGCCCTCGTTGCTGCGCATGCTGCGGAGCTGCCATCAGATGGAAGCACTCATTGTGTTGAGGTGATATTATGCTTGTCTGC
 AGTTGCTGCTGATGTTCTTCCCTGTAGAGACCAAGGGCCGTGGAATGGACTGATGCAATATCAGGCTTCAGCATCTAGATGAGCTGTAGT
 TCATCACTGTTTGTATGCTCATATTATGTACAGAGAATATACATGTAACATGGTGTGGATAGTTCCCTCAAATATGTG CAGGACTTTCCA
 AATTGTGAAGACAGATTTGTACTGAAAGACAGCTGGTCTGTACTGCCAAACACTGAAAGACAGAGTTGTATTACTCTCTCAGCATTI
CTTTCCAAGCCAGAACCTTTTGGCCAAATACAAGAGTTTGCAGGTACTCCAAAAA

下划直线部分为 RT-PCR 引物, 两条引物间的序列已经过测序验证; 黑框所示为起始密码和终止密码; 下划波浪线部分为电子克隆方法未能得到的序列; 小写字体的 aataca 为 mRNA 的加尾信号。

图 4 新基因 *MFSP* 的 cDNA 序列

3 结 论

1996 年, 美国的 Pioneer Hi-Bred International 和 Human Genome Sciences 公司合作开始了玉米 EST 计划研究; 1998 年, 由美国国家科学基金 (National Science Foundation) 资助启动了玉米基因发掘计划 (Maize Gene Discovery Project), 其工作内容之一就是 EST 测序和 cDNA 芯片的制作^[15]。随着大规模 EST 测序计划和功能基因组学研究的开展, 玉米的 EST 数据得以高速增加, 截止到 2007 年 12 月, NCBI 公共信息数据库中已有近 120 万条玉米 EST 序列, 我们可以充分利用这些核酸数据来克隆目的基因的全长序列。在本研究中, 我们以玉米基因 *PIS2* 的克隆为例显示了电子克隆技术的快速性和高效性。本实验室已经利用玉米 EST 数据库, 电子克隆并通过 RT-PCR 获得了多个感兴趣的基因, 进一步的功能验证工作正在进行中。

然而对于那些全长 cDNA 序列较长的基因来说, 由于绝大多数的 EST 仅是 cDNA 的 3' 端或 5' 端部分测序, 采用电子克隆的方法可能由于中间序列数据的缺少而无法直接得到全长序列。在这种情况下,

我们借用全基因组测序完成的水稻 DNA 数据库, 来克隆玉米中长基因的 cDNA 全长。我们以玉米干旱胁迫消减文库测序得到的一条 EST 为种子序列, 借助水稻的非冗余核酸数据库, 采用基于 EST 的电子克隆方法并结合 RT-PCR 技术成功地克隆到玉米基因 *MFSP* 的 cDNA 全长, 这一实验结果证实了借助水稻核酸数据库对玉米中较长基因全长序列克隆策略的可行性。我们又进一步利用玉米基因组数据库简便地克隆到 *MFSP* 的全长基因组序列, 为今后玉米基因 DNA 全长序列的克隆指出一个新思路。

参考文献

- [1] Wei M H, Karavanova I, Ivanov S V, et al. In silico-initiated cloning and molecular characterization of a novel human member of the L1 gene family of neural cell adhesion molecules. *Hum Genet*, 1998, 103(3): 355-364
- [2] Lander E S, Linton L M, Birren B, et al. Initial sequencing and analysis of the human genome. *Nature*, 2001, 409 (6822): 860-921
- [3] Kent W J. BLAT-the BLAST-like alignment tool. *Genome Res*, 2002, 12(4): 656-664
- [4] Peterson D G, Schulze S R, Sciara E B, et al. Integration of

- Cot analysis, DNA cloning, and high-throughput sequencing facilitates genome characterization and gene discovery. *Genome Res*, 2002, 12(5): 795-807
- [5] Lescure A, Daniel G, Philippe C, et al. Novel selenoproteins identified in silico and in vivo by using a conserved RNA structural motif. *J Biol Chem*, 1999, 274(53): 38147-38154
- [6] Huminiecki L, Bicknell R. In silico cloning of novel endothelial-specific genes. *Genome Res*, 2000, 10(11): 1796-1806
- [7] Chen Y, Zhao Y H, Wu R. In silico cloning of mouse Muc5b gene and upregulation of its expression in mouse asthma model. *Am J Respir Crit Care Med*, 2001, 164(6): 1059-1066
- [8] Kanehisa M, Bork P. Bioinformatics in the post-sequence era. *Nat Genet*, 2003, 33(Suppl): 305-310
- [9] The Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, 2000, 408(6814): 796-815
- [10] Yu J, Hu S, Wang J, et al. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *Indica*). *Science*, 2002, 296(5565): 19-92
- [11] Goff S A, Ricke D, Lan T H, et al. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *Japonica*). *Science*, 2002, 296(5565): 92-100
- [12] 王冬冬, 朱延明, 李勇等. 电子克隆技术及其在植物基因工程中的应用. 东北农业大学学报, 2006, 37(3): 403-408
- [13] 翟淑梅. 玉米磷脂酰肌醇代谢途径中酶的基因克隆及其功能鉴定: [博士毕业论文]. 济南: 山东大学生命科学学院, 2005
- [14] 岳桂东, 胡孝瑞, 庄云龙等. 干旱胁迫下开花期玉米消减文库的构建及分析. 山东大学学报, 2006, 41(4): 168-172
- [15] Whitelaw C A, Barbazuk W B, Pertea G, et al. Enrichment of gene-coding sequences in maize by genome filtration. *Science*, 2003, 302(5653): 2118-2120

Molecular cloning of maize *PIS2* and *MFSP* with bioinformatics databases

Yue Guidong, Sui Zhenhua, Li Xu, Zhang Juren

(School of Life Science, Shandong University, Jinan 250100)

Abstract

A cDNA containing the complete opening reading frame (ORF) of the putative *PIS2* (*phosphatidylinositol synthase* 2) gene was cloned from maize (*Zea mays* L.) by means of the RT-PCR method coupled with the *in silico* cloning approach based on the expressed sequence tags (EST) database (dbEST). Full-length cDNA of the maize *MFSP* (*major facilitator superfamily protein*) gene was isolated by the RT-PCR method combined with the *in silico* cloning with assistance of the *Oryza sativa* non-redundant nucleotide database. And by using the plant genome database (PlantGDB), the genomic sequence of the *MFSP* gene was obtained. These results of the study indicate that full-length cDNA and genomic sequences of genes can be easily obtained by making the best use of molecular bioinformatics databases.

Key words: *in silico* cloning, RT-PCR, maize (*Zea mays* L.)