

# 大豆吡哆醇生物合成蛋白基因 (*PDX*) 的电子克隆和进化分析

李蕊, 孟宪萍, 胡英考, 蔡民华, 李雅轩

(首都师范大学 生命科学学院, 北京 100037)

**摘要:** 电子克隆(*In silico* cloning) 是随着基因组计划和 EST 计划实施而发展起来的利用生物信息学手段进行基因克隆的新方法。根据物种间同源基因相对保守的特点, 以拟南芥(*Arabidopsis thaliana*) 吡哆醇生物合成蛋白 cDNA 序列为信息探针, 对大豆(*Glycine max*) EST 数据库进行同源搜索和序列拼接, 获得了 1 280 bp 长的大豆吡哆醇生物合成蛋白的基因序列(GenBank 登陆号为 DQ139265)。经过 RT-PCR 扩增、基因组 PCR 扩增、分子克隆和序列分析验证, 结果表明与电子克隆序列完全一致。该基因具有完整的开放阅读框架(ORF, 20~ 955 bp), 编码 311 个氨基酸。通过与水稻、日本百脉根、烟草、截形苜蓿等物种的吡哆醇生物合成蛋白序列比对, 发现该基因具有高度的保守性。表明根据物种间同源基因序列, 对跨物种间 EST 数据库进行同源检索、筛选、拼接, 是克隆基因的有效途径。

**关键词:** 电子克隆; EST 数据库; RT-PCR; 基因组 PCR; 大豆; 吡哆醇生物合成蛋白

中图分类号: S565.1 文献标识码: A 文章编号: 1000- 7091(2007)01- 0064- 05

## *In Silico* Cloning of Pyridoxine Biosynthesis Protein Gene from Soybean and Evolution Analysis

LI Rui, MENG Xian-ping, HU Ying-kao, CAI Min-hua, LI Ya-xuan

(College of Life Science, Capital Normal University, Beijing 100037, China)

**Abstract:** *In silico* cloning is developing with the project of genomic and EST. It is a new method of gene-cloning through bioinformatics. A query probe from the cDNA sequence of *Arabidopsis thaliana* pyridoxine biosynthesis protein was designed to blast *Glycine max* EST database according to the relative conservation of homologous genes and a soybean *PDX* gene cDNA of 1 280 bp was obtained (GenBank Accession, DQ139265). This sequence was confirmed by RT-PCR, genomic PCR, molecular cloned and sequenced. It contained a complete ORF, from 20 to 955 bp, encoded 311 amino acids, and conserved with *Arabidopsis thaliana*, *Oryza sativa*, *Lotus corniculatus* var. *japonicus*, *Nicotiana tabacum*, *Medicago truncatula* and *Triticum aestivum*. The results showed that it was an efficient technique to clone new genes by searching EST database with homologous gene of model species.

**Key words:** *In silico* cloning; EST database; RT-PCR; Genomic PCR; *Glycine max*; Pyridoxine biosynthesis protein

基于表达序列标签(Expressed sequence tags, EST)的电子克隆(*In silico* cloning)策略, 是近年发展起来的一门快速克隆基因的新技术。其技术核心是利用生物信息学技术组装延伸 EST 序列, 获得基因的部分乃至全长 cDNA 序列, 并利用实验的方法进行克隆分析、验证。随着 EST 数据库的进一步完善, 电子克隆策略已成为克隆新基因的重要方法, 并成功的应用于人类基因组的研究<sup>[1-4]</sup>。

吡哆醇生物合成蛋白(Pyridoxine biosynthesis protein)是维生素 B<sub>6</sub>合成过程中的一个限速酶。研

究表明, 吡哆醇生物合成蛋白主要与质膜和内膜系统有关, 推测维生素 B<sub>6</sub>对膜的功能有重要影响<sup>[5]</sup>。维生素 B<sub>6</sub>在体内的活性形式 5-磷酸吡哆醛(Pyridoxal 5-phosphate)(PLP), 它是许多代谢途径中酶的辅助因子, 也是植物体内的一个强抗氧化剂<sup>[6]</sup>。因此, 克隆编码吡哆醇生物合成蛋白的基因对研究维生素 B<sub>6</sub>的代谢过程, 进而进行转基因研究, 获得高维生素 B<sub>6</sub>以及具有强抗氧化性的植物品系具有重要意义。

本研究采用电子克隆的方法, 以拟南芥吡哆醇

收稿日期: 2006- 11- 02

基金项目: 北京市教委科技发展计划项目基金资助(200310028112)

作者简介: 李蕊(1982- ), 女, 北京人, 硕士, 主要从事分子遗传学方面的研究工作

通讯作者: 李雅轩(1965- ), 女, 北京人, 副教授, 主要从事分子遗传学方面的研究工作。

生物合成蛋白基因(*PDX*) cDNA 序列( GenBank 登录号为 AY972813) 为信息探针, 对大豆 EST 数据库进行 BLAST 检索和拼接, 获得了大豆的 *PDX* 基因的 cDNA 序列并采用 RT-PCR 方法和基因组 PCR 方法同时进行了验证。利用生物信息学的方法分析该 cDNA 序列的基因特征及所表达的蛋白质序列, 并与其他物种的同源基因表达的蛋白进行比较, 初步预测其功能。

## 1 材料和方法

### 1.1 生物信息学数据库及生物软件

生物信息学数据库: 美国国家生物信息技术中心(National Center for Biotechnology Information, NCBI) 提供并维护的 BLAST 数据库中的 nr, EST 数据库, 用 ORF Finder 预测拼接基因可能的开放阅读框。

生物软件: 用 BioEdit 软件进行 EST 片段的拼接; 用 DNAMAN 软件进行不同物种间同源基因的氨基酸序列比对及绘制系统发育树。

### 1.2 电子克隆的技术路线和方法

本研究采用生物信息学的方法, 根据不同物种间同源基因的核酸序列相对保守的特点, 成功克隆了大豆吡哆醇生物合成蛋白基因(*PDX*)。其技术路线为: 在 GenBank 的核酸(nr) 数据库中检索拟南芥的 *PDX* 基因序列, 获得拟南芥的 *PDX* 的基因序列 (AY972813), 以 AY972813 序列为探针对大豆 (*Glycine max*) EST 数据库进行 BLAST 检索, 获得与拟南芥 *PDX* 基因同源的大豆 EST 片段; 从获得的大豆 EST 片段中选出同源性最高的一条 BI941812, 再对大豆 EST 数据库进行 BLAST 检索, 得到多条与之高度同源的大豆 EST 序列; 将获得的 EST 序列进行首尾拼接, 获得新的 Contig; 以获得的 Contig 反复对大豆 EST 公共数据库进行 BLAST 检索、拼接, 尽可能获得全长 cDNA; 分析候选的基因特征, 如 ORF、推测编码的蛋白质序列, 并与其他物种同源基因进行比较, 初步预测其功能。

### 1.3 植物材料的培养

大豆种子于自来水中浸泡过夜, 将吸胀之后的种子种在花盆中, 光照培养室培养。待大豆出现真叶后取新鲜幼嫩的大豆叶片用于基因克隆。

### 1.4 大豆总 RNA 的抽提和 cDNA 第一链的合成

总 RNA 的抽提采用 Invitrogen 公司的 Trizol 试剂, 取 50~100 mg 大豆叶片于液氮中研磨成粉末, 加入 1 mL 的 Trizol, 研磨液冰上放置 10 min 后, 加入 350  $\mu$ L 的纯氯仿剧烈振荡后, 4  $^{\circ}$ C, 12 000 r/min 离心 10 min, 取上层水相于新的离心管中, 加等体积的异

丙醇混匀, -20  $^{\circ}$ C 沉淀 30 min, 4  $^{\circ}$ C, 12 000 r/min 离心 10 min, 回收 RNA, 弃上清, 按 1 mL Trizol 液至少 1 mL 的比例加入 75% 的乙醇洗涤沉淀, 4  $^{\circ}$ C, 12 000 r/min 离心 5 min, 弃上清, 室温或真空干燥 5~10 min, 用无 RNase 的灭菌水溶解沉淀。1% 琼脂糖凝胶电泳检测 RNA 的完整性。

1  $\mu$ g 总 RNA 用于 cDNA 第一链的合成, 合成步骤按 Reverse Transcription System 反转录试剂盒说明书进行, 试剂盒购自 Promega 公司。

### 1.5 大豆植株叶片基因组 DNA 的提取

采用 CTAB 法提取大豆植株叶片基因组 DNA。取大豆幼嫩叶片 1~2 g, 加入液氮研磨至粉末状, 加入 500  $\mu$ L CTAB 溶液于 65  $^{\circ}$ C 水浴中保温 1 h 后, 加入等体积的氯仿/异戊醇(24:1) 冰上放置 5 min, 10 000 r/min 离心 10 min, 取上清至另一新管中, 加入 2 倍体积的无水乙醇沉淀 10 min, 12 000 r/min 离心 15 min, 回收 DNA 沉淀, 用 70% 乙醇洗涤沉淀 2 次。稍干后, 用灭菌 ddH<sub>2</sub>O 溶解沉淀。1% 琼脂糖凝胶电泳检测 DNA 的完整性。

### 1.6 引物的设计与合成

根据电子克隆所得的 cDNA 序列, 分别在起始密码子区域和终止密码子区域设计特异引物(由上海生工生物工程公司合成)。

引物 1: 5'-CACTCTAGAATGGAAGGATCAGGATCCGGTG-3'

引物 2: 5'-ATAGAGCTCTCATTTCAGAGCGATTGCGAAG-3'

### 1.7 PCR 扩增大豆 *PDX* 基因片段

利用所设计的特异引物, 分别以大豆 cDNA 和基因组 DNA 为模板扩增特异片段。PCR 程序为: 94  $^{\circ}$ C 预变性 5 min, 进入循环 94  $^{\circ}$ C 变性 1 min, 67  $^{\circ}$ C 复性 1 min, 72  $^{\circ}$ C 延伸 1 min。共进行 35 个循环。72  $^{\circ}$ C 延伸 10 min。反应完毕后取 2  $\mu$ L 进行 1% 琼脂糖凝胶电泳分析。

### 1.8 分子克隆和 DNA 序列分析

用 1% 琼脂糖凝胶电泳纯化回收 PCR 产物, 连接到 pMD18-Tsimple 载体(摩尔比 3:1) 上, 挑选阳性克隆进行 PCR 及酶切鉴定, 委托 TaKaRa 公司进行 DNA 序列测定。

## 2 结果与分析

### 2.1 大豆 EST 库的搜索与序列拼接

利用拟南芥吡哆醇生物合成蛋白作为探针, 对大豆 EST 数据库进行 BLAST 检索, 得到一个高相似的 EST 序列 BI941812, 然后以此序列作为种子序列

BLAST 检索大豆 EST 数据库, 将检出的与种子序列同源性较高的或有部分重叠的 EST 序列(BI941812, BI942400, BM307795, BI892713) 拼接组装为重叠群 (Contig), 再以此重叠群序列重复以上 BLAST 检索过程, 反复进行 EST 重叠群序列的拼接和比对, 直至重叠群不能继续延伸。利用 BioEdit 软件进行拼接, 并用 DNAMAN 进行序列分析, 最终获得一个全长为 1 280 bp 的序列 (图 1)。通过 ORF Finder 软件预测, 其具有完整的开放阅读框 ORF (20~955 bp), 编码 311 个氨基酸。

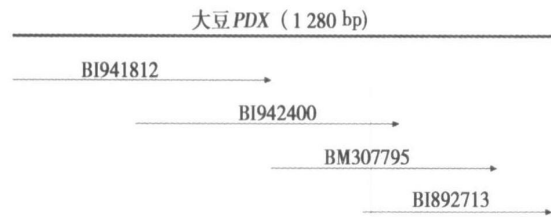


图 1 大豆 PDX 基因 cDNA 序列的拼接示意图

Fig 1 The joint map of *Glycine max* PDX cDNA sequence

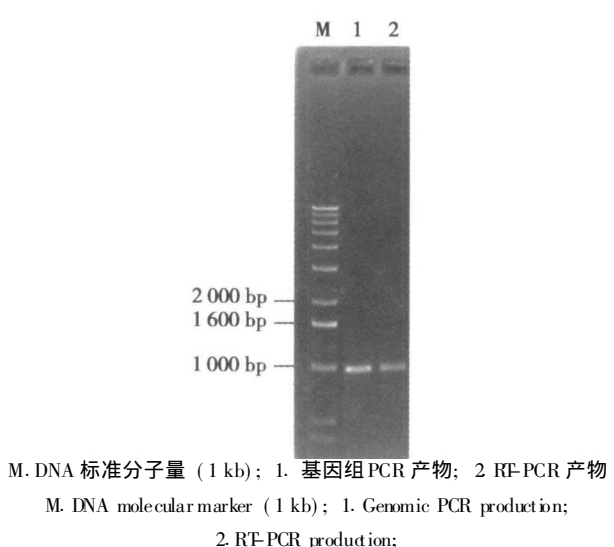


图 2 大豆 PDX 基因 RT-PCR 及基因组 PCR 产物的琼脂糖电泳分析

Fig 2 Identification of PCR production of *Glycine max* PDX gene by agrose gel electrophoresis

```

                                CCATTCGTCGGTGCCCTCA
PDX 20  ATGGAAGGATCAGGATCCGGTGTGGTTACAGTCTACGGCAACGGCGCCATCACCGAGACC
      M  E  G  S  G  S  G  V  V  T  V  Y  G  N  G  A  I  T  E  T
PDX 80  AAGAAGTCCCCCTTCTCCGTCAAAGTCGGCCTAGCCCAGATGCTCCGCGGCGGGCGTCATC
      K  K  S  P  F  S  V  K  V  G  L  A  Q  M  L  R  G  G  V  I
PDX 140 ATGGACGTCGTCGACGCCGAGCAGGCCGCGATCGCGGAGGAGGCCGCTGCCGCGGTG
      M  D  V  V  D  A  E  Q  A  R  I  A  E  E  A  G  A  C  A  V
PDX 200 ATGGCCCTGGAGCGTGTCCCGCCGACATCCGGGCCAGGGCGGGCTGGCCCGCATGAGC
      M  A  L  E  R  V  P  A  D  I  R  A  Q  G  G  V  A  R  M  S
PDX 260 GATCCGCAGCTCATCAAGGATATCAAGCGGGCCGTACGATCCCCGTGATGGCCAAGGCC
      D  P  Q  L  I  K  D  I  K  R  A  V  T  I  P  V  M  A  K  A
PDX 320 CGCATCGGCCACTTCGTGGAAGCCAGATCCTCGAAGCCATCGGCATCGACTACGTCGGC
      R  I  G  H  F  V  E  A  Q  I  L  E  A  I  G  I  D  Y  V  G
PDX 380 GAGAGCGAGGTCTCACCTCGCCGACGACGCCAACCACATCAACAAACACAACCTCCGC
      E  S  E  V  L  T  L  A  D  D  A  N  H  I  N  K  H  N  F  R
PDX 440 ATTCCTTCTGCTGTGGCTGCCGAACCTCGGCGAGGGCCCTCCGCGCATCCGCGAGGGC
      I  P  F  V  C  G  C  R  N  L  G  E  A  L  R  R  I  R  E  G
PDX 500 GCCGCCATGATCCGCACCAAGGGCGAGGCCGACACCGGCAACATCATCGAGGCCGTTTCG
      A  A  M  I  R  T  K  G  E  A  D  T  G  N  I  I  E  A  V  R
PDX 560 CACGTCCGCTCCGTTCATGAGCGACATTAGGGTTCTCCGCAACATGGACGACGACGAGGTC
      H  V  R  S  V  M  S  D  I  R  V  L  R  N  M  D  D  E  V
PDX 620 TTCACCTTCGCCAAGAACATCGCTGCGCCTTATGACCTCGTCATGCAGACCAAGCAGCTT
      F  T  F  A  K  N  I  A  A  P  Y  D  L  V  M  Q  T  K  Q  L
PDX 680 GGCGCGCTCCCGTTGTCCACTTCGCGCGCGGCGAGCGCCACCCCGCCGACGCGCGG
      G  R  L  P  V  V  H  F  A  A  G  G  A  A  T  P  A  D  R  E  A
PDX 740 CTCATGATGCAGCTCGGCTGCGACGGCGTCTTCGTCGGCTCCGGCGTTTTCAAGAGCGGC
      L  M  M  Q  L  G  C  D  G  V  F  V  G  S  G  V  F  K  S  G
PDX 800 GATCCCGCGAAGCGCGCGGCGATTGTTACGGCCGTGACTACTACAGCGACCCGTGAG
      D  P  A  K  R  A  R  A  I  V  Q  A  V  T  H  Y  S  D  P  E
PDX 860 ATTTTGGCGGAGGTTAGCTGCGGATTGGGTGAAGCCATGGTTGGTATCAATTTGACCGAT
      I  L  A  E  V  S  C  G  L  G  E  A  M  V  G  I  N  L  T  D
PDX 920 GACGAGGTTGAGAGGTTCCGCAATCGCTCTGAATGAAATGAATAATAATTCTCAGATTTG
      D  E  V  E  R  F  A  N  R  S  E  *
PDX 980 TTTCTCTGTGGCGCTTCAATTTGGACTATTAAGGGTTTTAGTTGCTTCTTATATATTATA
PDX 1040 ATAAGGTTATGTTTTAGTTTTCGGTGTTTTGAACCGTGTTAATTTTTTTTAGAAAAAC
PDX 1100 CAAAAAGAAAAAATAGATAGGTAGCTATTTGGGCATGCCTTGTAATATGCTATACTTGT
PDX 1160 TTTTACATATTACGCGTCGTGGTATTGTGTTGAGTTAGTAGTGCGTTCGTGGTAAACGAG
PDX 1220 AATCGAATCATTTCTGTTTTTGGATATTCAATCCTATTTTTGGGTTTAATTTTTTTTAA
PDX 1280 A
```

星号表示终止密码子 Stop codon is indicatde by a sterisk

图 3 大豆 PDX 的 cDNA 序列及推导的氨基酸序列

Fig 3 The nudeotide acid sequence and deduced amino acid sequence of *Glycine max* PDX

## 2.2 RT-PCR 验证

采用 RT-PCR 的方法对上述结果进行了验证, 经 1% 琼脂糖凝胶电泳分析, 结果在大约 936 bp 处

出现单一条带, 与预期的 PDX 基因的 ORF 区大小相符 (图 2)。用 PMD18-T simple 载体克隆后测序, 结果与电子克隆序列一致 (图 3)。运用 Sequin 软

件向 GenBank 递交该序列, 注册号为: DQ139265。

2.3 基因组 PCR 验证

经过网上比对分析, 发现拟南芥中 *PDX* 基因的 cDNA 序列与基因组 DNA 序列是完全一致的, 即该基因没有内含子, 并且该基因在各个物种中同源性非常高, 因此, 推测在大豆中 *PDX* 基因也可能没有内含子。本研究在采用 RF-PCR 方法验证的同时又运用了基因组 PCR 方法进行了验证, 证明

了我们的推测是正确的。基因组 PCR 结果经 1% 琼脂糖凝胶电泳检测, 在大约 936 bp 处出现单一条带, 与预期的 *PDX* 基因的 ORF 区大小相符。用 pMD18-Tsimple 载体克隆后测序, 结果与电子克隆序列以及 RF-PCR 结果完全一致。证明大豆 *PDX* 基因与拟南芥一样是没有内含子的, 与预测的结果是一致的。

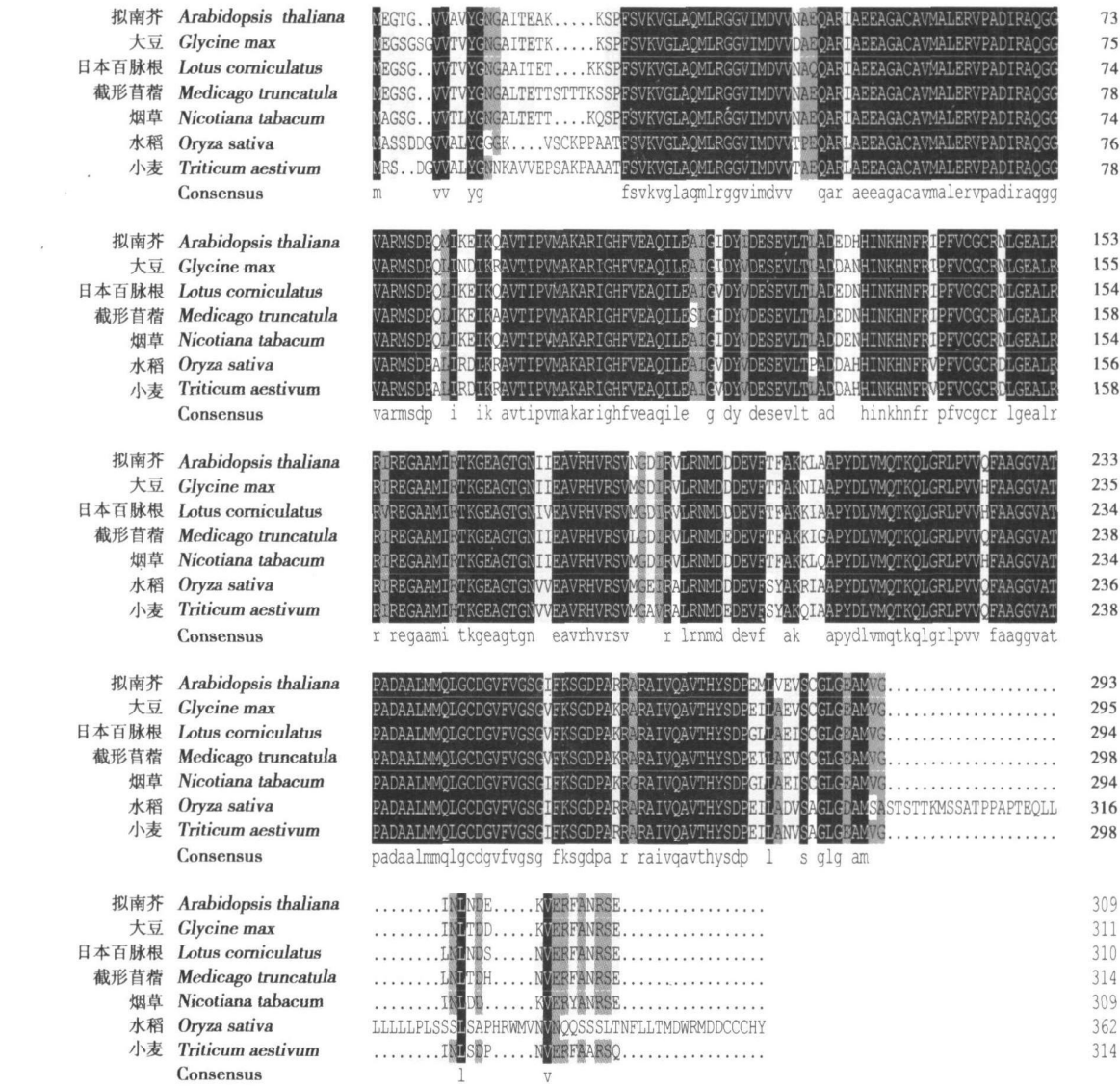


图 4 大豆 PDX 蛋白序列与其他植物同源蛋白序列的比较

Fig 4 An alignment of *Glycine max* PDX and homologous protein from other plants

2.4 不同物种间蛋白质同源性比较分析

将大豆的 *PDX* 基因与日本百脉根 (*Lotus corniculatus* var. *japonicus*)、烟草 (*Nicotiana tabacum*)、拟南芥 (*Arabidopsis thaliana*)、截形苜蓿 (*Medicago truncatula*)、小麦 (*Triticum aestivum*) 和水稻 (*Oryza sativa*) 等进行同源性比较发现, 该基因在蛋白水平上有着很

高的一致性, 分别为 92.95%, 92.31%, 91.96%, 91.77%, 84.81% 和 70.25% (图 4)。从 DNAMAN 软件构建的几种植物 *PDX* 系统发育树 (图 5) 可以看出, 双子叶植物大豆与拟南芥、截形苜蓿、日本百脉根和烟草的同源性较高, 单子叶植物水稻和小麦属于同一个分支, 它们与双子叶植物 *PDX* 蛋白的同源

性也很高, 证明 *PDX* 基因在植物中是十分保守的。系统发育树基本上代表了它们在经典分类上的地位。

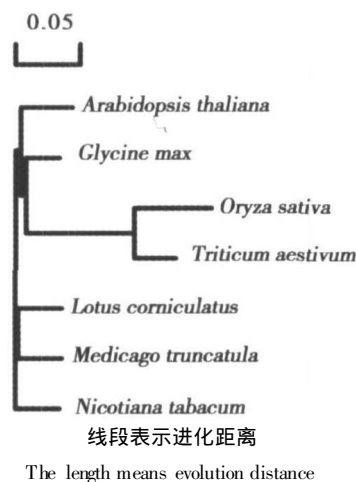


图5 植物 *PDX* 的系统发育树

Fig 5 Phylogenetic tree of plant *PDX*

### 3 讨论

利用 EST 进行电子克隆是克隆功能基因的新途径。与传统的基因克隆方法相比, 具有快捷、成本低、针对性强等特点, 但也受到 EST 数量和质量的限制<sup>[7]</sup>, 因此, 在 EST 资料非常丰富的模式物种(如人、小鼠等)中应用较多, 而植物基因的电子克隆还鲜有报道<sup>[8,9]</sup>。

维生素 B<sub>6</sub>是所有生物所必需的元素。它是许多代谢途径的辅酶因子, 并且最近研究发现, 它是一个强抗氧化剂<sup>[10]</sup>。*PDX* 基因是编码维生素 B<sub>6</sub>代谢途径中一个关键的限速酶——吡哆醇生物合成蛋白的基因, 在整个维生素 B<sub>6</sub>合成过程中起重要作用。研究表明, 植物和微生物可以进行体内的维生素 B<sub>6</sub>的合成, 而人类只能从食物中获得<sup>[11]</sup>。由此可见, 从大豆中成功克隆到 *PDX* 基因并实施转基因获得高表达的植物品种, 对人类具有重要意义。

通过对不同物种的 *PDX* 基因表达的蛋白质进行分析, 发现该基因编码的蛋白序列的同源性是非常高的。我们利用这一特性, 将已经研究获得的模式植物(拟南芥)的基因资源进行跨物种的 EST 数据库同源检索, 成功地克隆了目标物种的同源基因。

这充分证明了利用不同物种间关键基因的高度保守性, 采用现代生物信息学及实验验证相结合的技术方法, 是快速筛选重要功能基因的一种有效方法。

### 参考文献:

- [1] Zhang C, Yu Y, Zhang S, *et al.* Characterization, chromosomal assignment, and tissue expression of a novel human gene belonging to the ARF GAP family[J]. *Genomics*, 2000, 63(3): 400–408.
- [2] 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[J]. *Human Genetics*, 1998, 103(3): 355–364.
- [3] Hosoi T, Koguchi Y, Sugikawa E, *et al.* Identification of a novel human eicosanoid receptor coupled to G(i/o)[J]. *The Journal of Biological Chemistry*, 2002, 277(35): 31459–31465.
- [4] West A B, Lockhart P J, O'farrell C, *et al.* Identification of a novel gene linked to parkin via a bi-directional promoter[J]. *Journal of Molecular Biology*, 2003, 326(1): 11–19.
- [5] Chen H, Xiong L. Pyridoxine is required for post-embryonic root development and tolerance to osmotic and oxidative stresses[J]. *Plant J*, 2005, 44(3): 396–408.
- [6] Wang H, Liu D, Liu C, *et al.* The pyridoxal kinase gene *TaPdxK* from wheat complements vitamin B<sub>6</sub> synthesis-defective *Escherichia coli*[J]. *Journal of Plant Physiology*, 2004, 161: 1053–1060.
- [7] 陈大福, 牛宝龙, 翁宏飏, 等. 蜜蜂(*Apis mellifera*)腺苷酸转移载体基因 cDNA 的电子克隆[J]. *中国农业科学*, 2004, 37(6): 923–927.
- [8] 黄骥, 张红生, 曹雅君, 等. 一个新的 C<sub>2</sub>H<sub>2</sub> 型水稻锌指蛋白 cDNA 的克隆与序列分析[J]. *南京农业大学学报*, 2002, 25(2): 110–112.
- [9] 林慧贤, 刘筱斌, 李发强. 一个水稻小 GTP 结合蛋白基因 *Osrab5B* 的克隆和分析[J]. *高技术通讯*, 2001, 11(3): 9–14.
- [10] Ehrenshaft M, Bilski P, Li M, *et al.* A highly conserved sequence is a novel gene involved in de novo vitamin B<sub>6</sub> biosynthesis[J]. *Microbiology*, 1999, 96: 9374–9378.
- [11] Tambascio-Studart M, Titiz O, Raschle T, *et al.* Vitamin B<sub>6</sub> biosynthesis in higher plants[J]. *PNAS*, 2005, 102(38): 13687–13692.