生 物 化 学 与 生 物 物 理 学 报
ISSN 0582-9879 ACTA BIOCHIMICA et BIOPHYSICA SINICA 2001, 33(2): 163 - 172 CN 31-1300/Q

# 对水稻第 4 号染色体长臂近端粒区一个过氧化物酶基因簇的结构分析

陈泽华 周 波 韩 斌 钱跃民 洪国藩*
(中国科学院国家基因研究中心, 上海 200233)

**摘要** 通过对定位 BAC 克隆 q3037(H0207F01)的序列测定和分析, 在其中一个 22.5 kb 的区域发现一个由 5 个第三类过氧化物酶基因(依次命名为 *osp*1、*osp*2、*osp*3、*osp*4、*osp*5)组成的基因簇。分析表明, *osp*1、*osp*2 和 *osp*3 分别含 1 个内含子, *osp*4 和 *osp*5 分别含 2 个内含子。该 5 个基因分别编码 338、335、336、343 和 346 个氨基酸残基的蛋白质, 而且都具有 N 端信号肽序列, 其中 OSP1、OSP4、OSP5 为阴离子过氧化物酶, OSP2、OSP3 为阳离子过氧化物酶。对 5 个基因间的两两比较分析和进化分析结果表明: 该基因簇是通过一系列的串联基因复制事件而形成; *osp*5 与来自玉米的 *ap*1 和来自大麦(*Hordeum vulgare*)的 *prx*7 为潜在的直向同源基因, 而且, *osp*1-5 与 *ap*1、*prx*7 构成了分泌性植物过氧化物酶基因家族中一个新的分枝。

**关键词** 过氧化物酶; 基因簇; 串联基因复制; 直向同源基因

含血红素的过氧化物酶(EC1.11.1.7)广泛分布于动、植物和细菌、真菌中, 它能氧化一系列的有机和无机物, 同时消耗 $H_2O_2$[1]。根据序列相似性, 来源于细菌、真菌和植物的过氧化物酶可归为同一个超家族, 与动物的过氧化物酶家族起源不同, 差异显著。来源于细菌、真菌和植物的过氧化物酶基因家族又可分为三类: 胞内过氧化物酶(I 类)、分泌性真菌过氧化物酶(II 类)和分泌性植物过氧化物酶(III 类)[2]。植物过氧化物酶被认为与植物的生长、发育、逆境耐受、防御反应等有关, 直接参与了细胞壁的木质化和角质化、生长素代谢、受伤组织的栓化愈合和对病原的防御反应[2]。每一种植物拥有近 100 个第三类过氧化物酶[3], 由此形成一个庞大的植物过氧化物酶基因家族。第三类过氧化物酶又可根据其等电点分为中性过氧化物酶、阴离子过氧化物酶和阳离子过氧化物酶。多数的第三类过氧化物酶基因具有 4 个外显子、3 个内含子, 而且内含子插入的位点严格保守[3]。

基因家族的不同成员在基因组中有几种可能的排布方式: 或者弥散分布于整个基因组, 或者在染色体的特定区段成簇分布, 或者两种方式兼而有之。家族基因的成簇分布现象在很多基因组中都被观察到过, 并被广泛研究。例如, 3 个水稻的 α-淀粉酶基因被发现成簇分布于一个 28 kb 的区域[4]; 对光敏感的、编码二磷酸核酮糖羧化酶小亚基的基因家族在单子叶、双子叶植物中都存在成簇分布现象[5]; 在拟南芥菜基因组中, 3 个腈水解酶基因(*nit*2/ *nit*1/ *nit*3)成簇分布于 13.8 kb 的区域[6], 5 个细胞壁偶联的受体激酶基因(*wak*1～5)集中分布在一个 30 kb 的区域[7]。

在基因家族进化过程中, 普遍存在有基因复制现象[8]。而复制后的基因位点就有可能进一步复制而形成多基因簇。在基因簇中只要有一个基因还保留着最初基因的功能, 那么其他成员的功能有可能发生较大的分化, 而组织特异性表达或发育时间特异性表达也随之发生变化。

植物过氧化物酶家族非常庞大, 但是对于家族成员在基因组中的分布情况却知道不多。在 *Trametes versicolor* 基因组中, 两个木质素过氧化物酶基因和一个锰离子过氧化物酶基因成簇分布于 10 kb 的区域[9]。在 *Populus kitakamiensis* 基因组中, 两个阴离子过氧化物酶基因分布于一个 7 kb 的区域[10]。在水稻基因组的遗传图构建和大规模 EST 测序过程中, 26 个过氧化物酶位点, 114 个过氧化物酶基因得到鉴定[11], 但基因在染色体上的详细分布并不知道。其中有一个过氧化物酶位点

收稿日期: 2000-11-08 接受日期: 2000-12-21

* 联系人: Tel, 021-64516371; Fax, 021-64825775; e-mail, gfhong@newnetra.ncgr.ac.cn

(R2184S)位于第4号染色体长臂近端粒区。在以前的工作中,利用R2184S作为探针将BAC克隆重叠群816定位到该区域[12]。从重叠群816中选择了BAC q3037(H0207F01)进行全序列测定并进行序列分析,发现了一个由5个过氧化物酶基因组成的基因簇。本文通过对该基因簇的详尽分析,揭示了基因之间的演化关系。

## 1 材料和方法(Materials and Methods)

### 1.1 序列测定

采用双末端测序战略(pairwise end sequencing)[13]对水稻BAC q3037进行了全序列测定。大致的过程为:首先用超声波将纯化的BAC DNA打碎,然后用绿豆核酸酶对打碎的DNA片段进行末端修补,修补完毕回收2～4 kb的片段,回收的片段再与经*Sma*I酶切并脱磷处理的pKS载体连接,连接产物转化DH5α,然后随机挑选亚克隆进行双末端测序,对所得序列进行组装(Phred/Phrap)[14,15],最后再用引物定向步行的方法填补余下的少数缺口。组装的正确性可从亚克隆框架和*Not*I酶切分析得到保证。

### 1.2 序列分析

利用NCBI(http://www.ncbi.nlm.nih.gov/BLAST)的BLAST程序进行数据库检索[16]。GCG软件包(10.1版本)被用来进行DNA和蛋白质序列的分析。运用Testcode[17]和Genscan[18]进行潜在的蛋白质编码区的预测和基因结构预测。多序列对准由PileUp程序进行。信号肽预测由SPScan程序进行。蛋白质的分子量和等电点计算由Compute pI/Mw(http://www.expasy.ch/tools/pi_tool.html)程序进行。蛋白质氨基酸序列两两比较由Gap程序进行。系统进化树的构建由GrowTree程序完成。

## 2 结果(Results)

### 2.1 序列测定

完成以后的q3037全序列全长124.3 kb(GenBank登录号AJ245900),平均序列丰度为10.5。经亚克隆框架分析和*Not*I酶切分析,确保了组装的正确性。同时,高丰度的序列覆盖也确保了最终完成序列的质量。在对该序列的初步分析时发现,其中22.5 kb的区域包含一个过氧化物酶基因簇,然后就对该区域进行了详尽的分析。

### 2.2 基因鉴定

利用GCG软件包(10.1版本)的TestCode程序对该22.5 kb区域进行了分析,结果表明,5个蛋白质编码区不均匀地分布于该区域[图1(A)]。将相应的蛋白质编码区分别进行BLASTx分析,显示都与数据库中的植物过氧化物酶有显著相似,表明该22.5 kb区域包含一个由5个过氧化物酶基因组成的基因簇。这5个基因依次命名为*osp*1、*osp*2、*osp*3、*osp*4和*osp*5,其中*osp*5对应于遗传标记R2184S。该5个基因具有相同的转录方向,基因间隔区分别为2、5.5、1.7和3.3 kb。

### 2.3 基因结构预测

对该22.5 kb序列运用Genscan软件分析,预测了5个过氧化物酶基因的结构。结果表明,*osp*4和*osp*5分别含3个外显子和两个内含子,而*osp*1,*osp*2和*osp*3只有两个外显子,被单一的内含子所切断[图1(B)]。

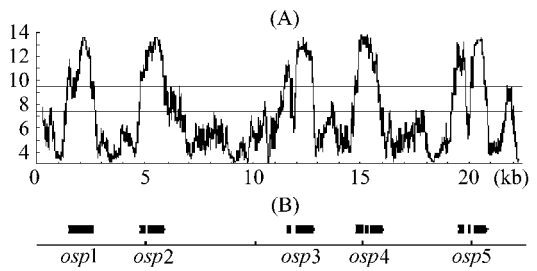


**Fig. 1 Gene prediction in the 22.5 kb region**

(A) TestCode analysis of the 22.5 kb sequence. The top region is supposed to predict coding regions to a 95% level of confidence. The bottom region is supposed to predict non-coding regions to the same confidence level[17]. (B) Genscan prediction of the 22.5 kb sequence. Gene structures are predicted using the matrix of *maize* (http://genes.mit.edu/GENSCAN.html)[18].

### 2.4 数据库检索

将5个过氧化物酶基因编码区分别用于数据库检索,当EST全长的90%以上与待检索序列的相同率大于95%以上时被认为是相互匹配的[19]。在对核酸数据库(nr)进行检索时,得到一个与*osp*5匹配(相同98.8%)的全长cDNA序列*oscpx*1(GenBank登录号:AF019743)。在对EST数据库(dbEST)检索时,得到来源于不同组织的分别与*osp*1、*osp*3、*osp*5相匹配的若干EST条目(表1),这些EST一部分与基因的5′端相匹配,一部分与基因的3′端相匹配。而*osp*2,*osp*4则没有相匹配的EST或cDNA序列。

将*osp*1、*osp*3、*osp*5分别与相匹配的cDNA或EST序列对准,可准确确定基因的内含子剪切位点。而*osp*2、*osp*4与相似程度较高的cDNA或EST进行对准,也能确定出基因的内含子/外显子边界

(图 2)。所有 5 个基因的内含子中,除 *osp*4 的 2 个内含子, *osp*5 的第二个内含子遵循 GT/AG 的序列模式以外,其余内含子都为 GC/AG 的序列模式。对 5 个基因的外显子和内含子计算 GC 含量发现,5 个基因的外显子 GC 含量在 64.8 %～69.4 %,而内含子 GC 含量在 25 %～45.5 %(表 2)。内含子的高 AT 含量保证了在基因转录后加工时的有效剪切[20]。

```
osp1      436  gccacaagctagcacagcacagccggtaagagagtgagtgcggagagccaaattgagatggtgcagcgagcaggagact 515
AU075454  1        caagctagcacagcacagccggtaagagagtgagtgcggagagccaaattgagatgggtgcagcgagcaggagact 76
AU031972  12   gccacaagctagcacagcacagccggtaagagagtgagtgcggagagccaaattgagatgggtgcagcgagcaggagact 91

osp1      516  ggccgtcctggagctcgtctccatcgtcgccgtcctcctcatctcctcaccagcagcagcagcagagttgtccgtggact 595
AU075454  77   ggccgtcctggagcggntctccatcgtcgccgtcctcctcatctcctcaccagcagcagcagcagagttgtccgtggact 156
AU031972  92   ggccgtnctggagnnnccctacatcgtcgccgtcctcctcatctcctcaccagcagcagcagcagagttgtccgtggact 171

osp1      596  tccacgcggcgtcgtgcccgcagctggagtccatcgtgcgctcctccgtgcaagccgctctccagcaggagatcgccctc 675
AU075454  157  tccacgcggcgtcgtgcccgcagctggagtccatcgtgcgctcctccgtgcaagccgctctccagcaggagatcgccctc 236
AU031972  182  tccacgcggcgtcgtgcccgcagctggagtccatcgtgcgctcctccgtgcaagccgctctccagcaggagatcgccctc 251

osp1      676  gccgccggcctcctccgcatcttcttccacgactgcttcccgcaggcacgcacgcaagaacatttgcacgtaccgacgaa 755
AU075454  237  gccgccggcctcctccgcatcttcttccacgactgcttcccgcag                                   281
AU031972  252  gccgccggcctcctccgcatcttcttncacgactgcttcccgcag                                   296

osp1      756  caatgctgtactgaactgaacatacatatatatatgcaggggtgcgatgcgtcggtgtatctgagaggaggcagcaactc 835
AU075454  282                                        ggtngcgatgcgtcggtgtatctgagaggaggcagcaactc 322
AU031972  297                                        ggttgcgatgcgtcggtgtatctgagaggaggcagcaactc 337

osp1      836  ggagcagggcatgggcccaaacc-tcacgctgcagccgc-gggcgctgcagctggtggaggacatccgcgccaaggtgca 913
AU075454  323  ggagcagggcatgggcccaaacc-tcacgctgcagccgcggggcgctgcagctggtggaggacat                 386
AU031972  338  ggagcagggcatgggcccaaaccttcacgctgcagccgc-gggcgctgcagctggtggaggacatccgagccaaggtgca 416

osp1      914  cgccgcctgcggccccaccgtctcctgcgccgacatctccgccctcgccacccgcgacgctgtcgtcgtctccggcggcc 993
osp1      994  cctcttacgccgtgccactcggccaaaaagacagcctcgctccggcatcgttggacctcgtcggagatctcccgggcccg 1073
osp1      1074 ggcacctccagagtccaggatctcatcgacctttttgcaagcaggggactgcgggacgccgccgacctggtggcgctctc 1153
AU030963  1            agagtccag-anctcatcgacctttttgcaaccaggggantgcgggacgccgccgacctggtggcgttctc 70

osp1      1154 cggcgggcacaccgtcgggaggacacgctgcgccttcttcgatgaccgcgcccgccgccaggacgacaccttctccaaga 1233
AU030963  71   cggcgggcacaccgtcgggaggacacg-tgcgccttttcgatgaccgcgcccgccgccaggacgacaccttctccaaga 149

osp1      1234 agctggcgttgaactgcaccaaggacccaaccggctgcagaacctggacgtgatcaccccgacgccttcgacaacgcc 1313
AU030963  150  agctggcgttgaactgcaccaaggacccaaccggctgcagaacctggacgtgatcaccccgacgccttcgacaacgcc 229

osp1      1314 tactacatcgcgctcacccacaaccagggcgtcttcacctccgacatggcgctcatcaaggaccggatcacggcgcccat 1393
AU030963  230  tactacatcgcgctcatccacaaccagggcgtcttcacctccgacatggcgctcatcaaggaccggatcacggcgcccat 309

osp1      1394 cgtcaggcagttcgccacggacaaggccgccttcttcacgcagttcgccaagtccatggtcaaactaagcaatgtgccga 1473
AU030963  310  cgtcaggcagttcgccacggacaaggccgccttcttcacgcagttcgccaagtccatggtcaaactaagcaatgtgccga 389

osp1      1474 ggactgacaggaacgtcggcgagatccgccgcagctgcttcaggaccaacagccagagcctcgtcgacttcgccaccagc 1553
AU030963  390  ggactgacaggaacgtcggcgagatccgccgcagctgcttcaggaccaacagccagagcctcgtcgacttcgccaccagc 469

osp1      1554 gatgaggagggcttcgctgcttctgcttgatcattcagctgtctgtcgttgctcgttaattagggagctaggtgcttgac 1633
AU030963  470  gatgagga-ggnttcgctgcttctgcttgatcattcagctgtctgtcgttgctcgttaattagggagctaggtgcttgac 549

osp1      1634 gtcgctagctgatcctagtctctttagtcttcaggccttggcttttgtacgcatatgctttcgttatgtccgtgcaataa 1713
AU030963  550  gtcgctagctgatcctagtctctttagtcttcaggccttggcttttgtacgcatatgctttcgttatgtccgtgcaataa 629
```

**Fig. 2A　The genomic sequence of** osp**1 and its alignment with matched ESTs**

The translational start (5′AGT) and stop codons (5′TGA or 5′TAA) are highlighted and underlined. The introns are underlined and in italic.

```
osp2      277  agagtgagagcagagagccaaattaagatggctgcagcgaggagactgcccgtcctggagctggtctccttcgtcgccgt 356
AU075454  27   agagtgagtgcggagagccaaattgagatgggtgcagcgaggagactggccgtcctggagcggntctccatcgtcgccgt 106
AU031972  42   agagtgagtgcggagagccaaattgagatgggtgcagcgaggagactggccgtnctggagnnnccctacatcgtcgccgt 121

osp2      337  cttccttatctcctcaccaacagcagcagcagccgagttgtctgtggacttccacgcggcgtcgtgcccgtcgctggagg 436
AU075454  107  cctcctcatctcctcacca---gcagcagcagcagagttgtccgtggacttccacgcggcgtcgtgcccgcagctggagt 186
AU031972  122  cctcctcatctcctcacca---gcagcagcagcagagttgtccgtggacttccacgcggcgtcgtgcccgcagctggagt 201

osp2      397  ccatcgtgagctcctccgtgcaagccgcgctgcagcaagagatcgccctcgccgccggcctcctccgcatcttcttccac 516
AU075454  147  ccatcgtgcgctcctccgtgcaagccgctctccagcaggagatcgccctcgccgccggcctcctccgcatcttcttccac 266
AU031972  162  ccatcgtgcgctcctccgtgcaagccgctctccagcaggagatcgccctcgccgccggcctcctccgcatcttcttncac 281

osp2      517  gactgcctcccgcaggcacgcacgcaagaacatactcctatattcgcacgtacgtacgaacaacaatgttgtactcaact 596
AU075454  267  gactgcttcccgcag                                                               281
AU031972  282  gactgcttcccgcag                                                               296

osp2      577  gaacatgctgcatatgcatgtgcaggggtgcgatgcatcggtgtatcttaggggaggcagcaactcagagcagggcatgg 676
AU075454  282                          ggtngcgatgcgtcggtgtatctgagaggaggcagcaactcggagcagggcatgg 336
AU031972  297                          ggttgcgatgcgtcggtgtatctgagaggaggcagcaactcggagcagggcatgg 351

osp2      637  gcccaaacctgacgctgcagccgcgggcgctgcagctcgtcgacgacatccgcgccagggtgcatgcagcatgcggcccc 756
AU075454  297  gcccaaacctcacgctgcagccgcgggcgctgcagct                                         374
AU031972  312  gcccaaacctcacgctgcagccgcgggcgctgcagctggtggaggacatccgagccaaggtgca               416

osp2      757  accgtctcctgcgccgacatctccgcgctcgccacccgtgacgccgtcgtcgtctccgggggaccctcgtacgcagtgtc 836
osp2      817  tctcggccaaaaagacagcctcgccccggccccggtgcgcctcgtcaaccaactcccgggcccgggcacctccagcgtcc 916
osp2      877  aggctctcctcgacaaattcggaagcaagggcctgagagaggccgccgacctggtggcgctctccggcgcccacaccgtc 996
AU030963  48                                          gccgccgacctggtggcgttctccggcgggcacaccgtc 86

osp2      997  gggagggcgcactgcgacttcttccgtgaccgcgccgcccgccaggacgacaccttctccaagaagctggccgtgaactg 1076
AU030963  87   gggaggacacg-tgcgccttttcgatgaccgcgcccgccgccaggacgacaccttctccaagaagctggcgttgaactg 165

osp2      1057 caccaaggacccaaccggctgcagaatctggacgtcgtcacgcccgacgccttcgacaacgcctactacgtcgcgctca 1156
AU030963  146  caccaaggacccaaccggctgcagaacctggacgtgatcaccccgacgccttcgacaacgcctactacatcgcgctca 245

osp2      1117 ccaggaagcaggggtgttcacctccgacatggcgctcatcaaggaccggatcacggcgcccaatcgtcaggcagttcgcc 1236
AU030963  206  tccacaaccagggcgtcttcacctccgacatggcgctcatcaaggaccggatcacggcgcccatcgtcaggcagttcgcc 325

osp2      1237 gcggacaaggccgcattcttcacgcagttcgccaagtccatggtcaagctcagccaggttccgaggactgacaggaacgt 1336
AU030963  326  acggacaaggccgccttcttcacgcagttcgccaagtccatggtcaaactaagcaatgtgccgaggactgacaggaacgt 425

osp2      1297 cggcgagatccgccgcagctgcttcaggaccaacggaccacgcctcgtcgacttggccaccggcgatgaggccgcttctc 1396
AU030963  386  cggcgagatccgccgcagctgcttcaggaccaacagccagagcctcgtcgacttcgccaccagcgatgagg          476
```

**Fig. 2B　The genomic sequence of** osp**2 and its alignment with high similar ESTs**

The translational start (5′ATG) and stop codons (5′TGA or 5′TAA) are highlighted and underlined. The introns are underlined and in italic.

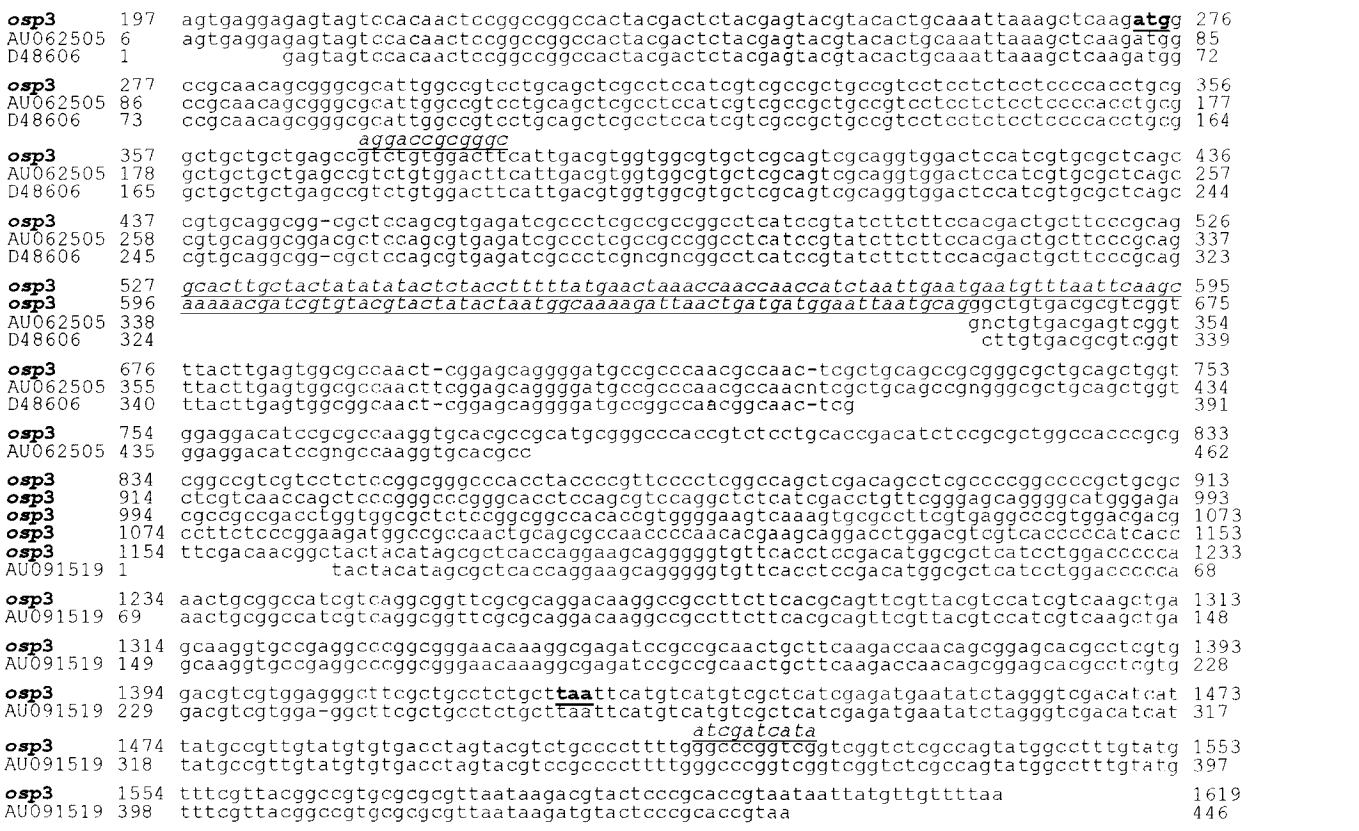

**Fig. 2C The genomic sequence of** osp**3 and its alignment with matched ESTs**

The translational start (5′AGT) and stop codons (5′TGA or 5′TAA) are highlighted and underlined. The introns are underlined and in italic.

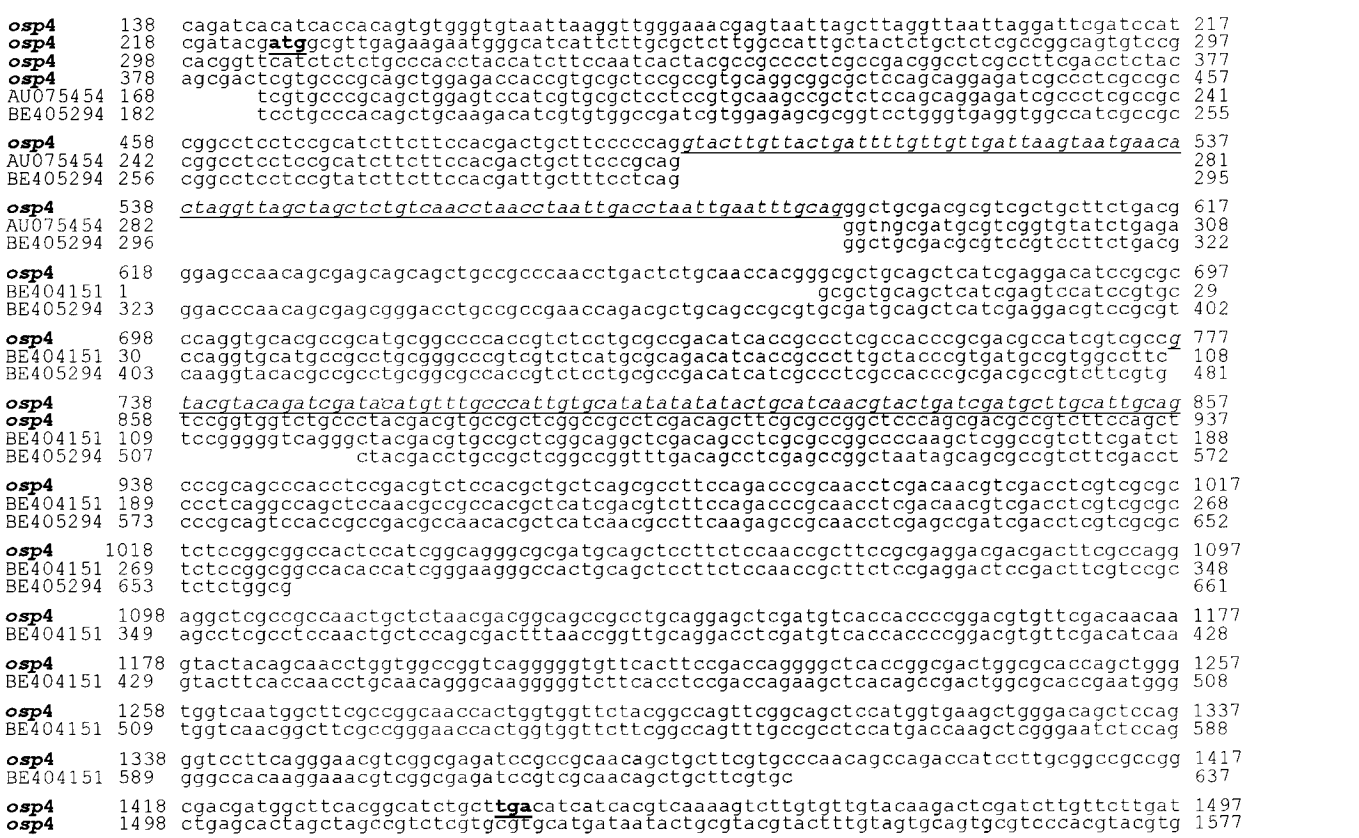

**Fig. 2D The genomic sequence of** osp**4 and its alignment with high similar ESTs**

The translational start (5′ATG) and stop codons (5′TGA or 5′TAA) are highlighted and underlined. The introns are underlined and in italic.

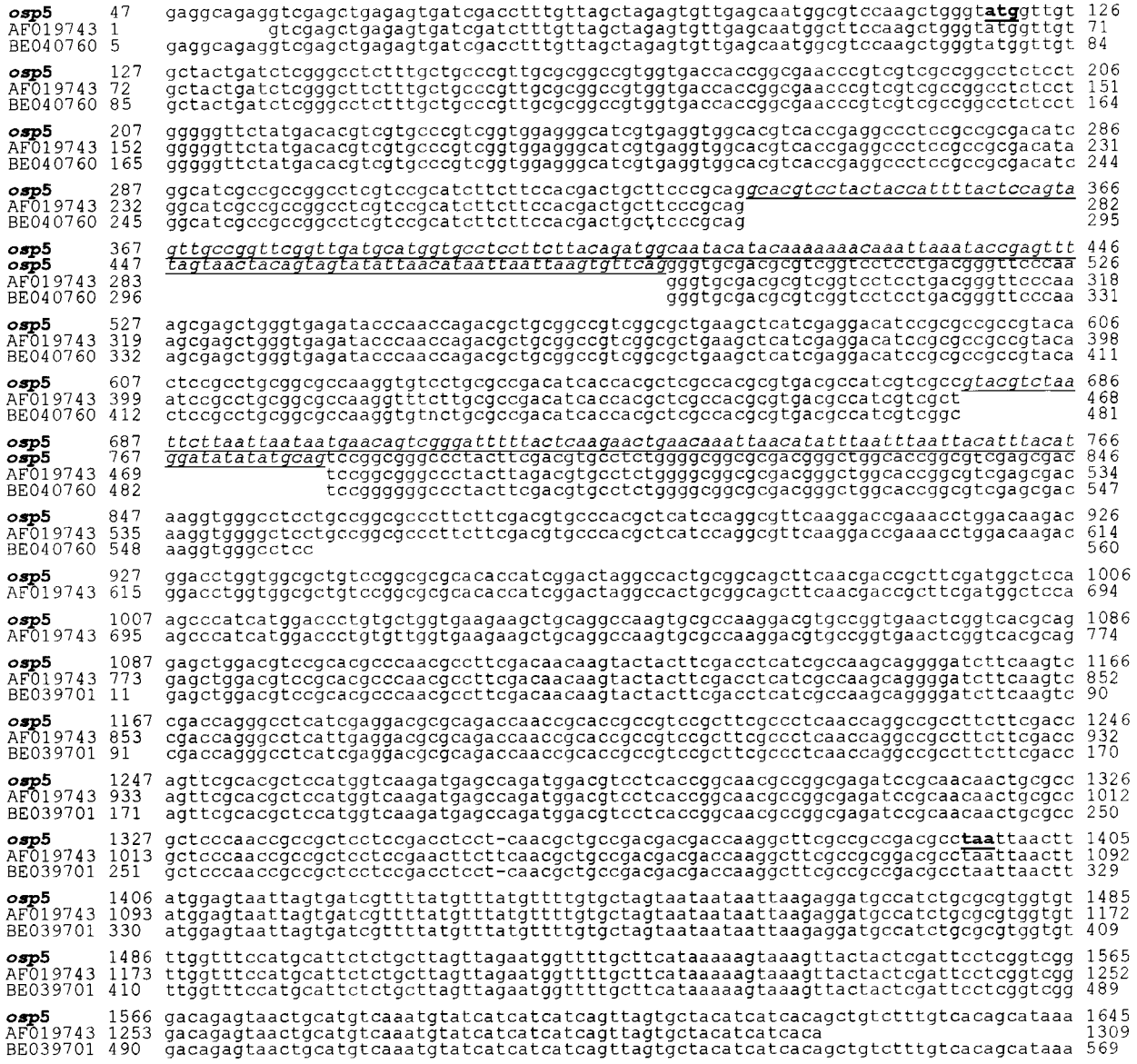


**Fig. 2E The genomic sequence of** osp**5 and its alignment with matched ESTs**

The translational start (5′ATG) and stop codons (5′TGA or 5′TAA) are highlighted and underlined. The introns are underlined and in italic.

### 2.5 预测的过氧化物酶结构

5 个基因分别编码 338、335、336、343、346 个氨基酸残基的蛋白质。典型的第三类过氧化物酶大小通常在 330～360 个氨基酸。在 5 个过氧化物酶的 N 端预测了信号肽序列,该信号肽可负责将相应的过氧化物酶导向细胞外。根据氨基酸序列计算 5 个过氧化物酶的分子量分别为: 33.3、32.7、32.1、34.7、34.8 kD。而且根据计算,OSP1、OSP4、OSP5 的等电点分别为 5.75、4.69、5.42,因此属于阴离子过氧化物酶; 而 OSP2、OSP3 的等电点分别为: 8.17、8.18,因此属于阳离子过氧化物酶(表 3)。这表明阳离子过氧化物酶基因和阴离子过氧化物酶基因可在染色体上紧密连锁。

通过多序列对准可发现在过氧化物酶家族中最为保守的三个结构域[图 3(B),(D),(F)],其中(B)、(F)为血红素结合区,而(D)为功能未知的保守区[21]。对于三类过氧化物酶都保守的 9 个氨基酸残基中[2],除了在 OSP1-3 中 123 位的 Arg(R)被 Gln(Q)所替代以外,其余 8 个残基在 OSP1～5 中都被观察到(图 3)。绝大多数第三类过氧化物酶保守的残基都在 OSP1～5、AP1、PRX7 中发现。第三类过氧化物酶特异的二硫键形成残基也被观察到: $Cys^{11}$-$Cys^{91}$、$Cys^{44}$-$Cys^{49}$、$Cys^{97}$-$Cys^{301}$、$Cys^{177}$-$Cys^{209}$(以成熟的辣根过氧化物酶 HRP C1 为参照)[2],但 OSP3 的 $Cys^{11}$前移了两位。除此以外,OSP1～5、AP1 和 PRX7 又表现出一些不同于其他亚家族的新

**Fig. 3 Alignment and comparison of the clustered peroxidases OSP1-5 and other peroxidases from class Ⅲ peroxidase family**

The amino acid sequences are given in one-letter code and have been aligned by introduction of gaps to maximize similarities. Predicted signal peptides are in italic. Horizontal lines represent conserved catalytic and distal heme binding domain (B), proximal heme binding domain(F), and central conserved domain (D) of unknow function. Conserved residues in class Ⅲ peroxidase are highlighted and indicated by *. Conserved cysteines(C)[2] involved in formation of disulfide bridges in class Ⅲ peroxidses are highlighted and indicated by ♯. Conserved residues in OSP1-5, AP1 and PRX7 and highlighted.

**Table 1 Matched ESTs( > 95 % identity) for** osp**1**, osp**3 and** osp**5**

| ACC. No. | Gene | Tissue | cDNA clone | Score |
|---|---|---|---|---|
| AU075805 | *osp*1 | Immature leaf | E60486 | 531 |
| AU030963 | *osp*1 | Immature leaf | E60486 | 1168 |
| D25030 | *osp*1 | Root | R2957 | 529 |
| AU031972 | *osp*1 | Root | R2957 | 513 |
| AU075454 | *osp*1 | Immature leaf | E60731 | 537 |
| AU031073 | *osp*1 | Immature leaf | E60731 | 682 |
| BE039997 | *osp*1 | Root | | 519 |
| AU091519 | *osp*3 | Callus | C11468 | 785 |
| AU062505 | *osp*3 | Callus | C11468 | 446 |
| D48606 | *osp*3 | Green shoot | S14924 | 446 |
| C20540 | *osp*3 | Green shoot | S14924 | 753 |
| D41595 | *osp*3 | Etiolated shoot | S4191 | 329 |
| C20517 | *osp*3 | Etiolated shoot | S4191 | 595 |
| AU031487 | *osp*3 | Immature leaf | E61710 | 454 |
| BE229129 | *osp*3 | Immature seed | 98BS0186 | 729 |
| BE229186 | *osp*3 | Immature seed | 98BS0275 | 676 |
| BE229213 | *osp*3 | Immature seed | 98BS0313 | 410 |
| AF019743 | *osp*5 | | | 1602 |
| D24571 | *osp*5 | Root | R2184 | 486 |
| AU031855 | *osp*5 | Root | R2184 | 636 |
| C20483 | *osp*5 | Root | R0894 | 759 |
| D24028 | *osp*5 | Root | R0894 | 569 |
| D20490 | *osp*5 | Root | R1777 | 559 |
| C20491 | *osp*5 | Root | R1778 | 700 |
| BE039701 | *osp*5 | Root | | 1134 |
| BE040760 | *osp*5 | Entire plant | | 579 |
| BE039233 | *osp*5 | Root | | 551 |
| AW155066 | *osp*5 | | mgie0001B21f | 537 |
| BE607327 | *osp*5 | Root | | 452 |

**Table 2 GC content of the introns/exons of the clustered genes,** osp**1-5**

| Gene | | Length(bp) | GC content(%) |
|---|---|---|---|
| *osp*1 | Exon1 | 228 | 66.7 |
| | Intron | 74 | 43.2 |
| | Exon2 | 789 | 65.8 |
| *osp*2 | Exon1 | 228 | 65.8 |
| | Intron | 90 | 45.5 |
| | Exon2 | 780 | 66.7 |
| *osp*3 | Exon1 | 243 | 67.9 |
| | Intron | 143 | 32.2 |
| | Exon2 | 768 | 67.9 |
| *osp*4 | Exon1 | 273 | 64.8 |
| | Intron1 | 93 | 34.4 |
| | exon2 | 186 | 71.5 |
| | Intron2 | 81 | 40.7 |
| | Exon3 | 588 | 66.8 |
| *osp*5 | Exon1 | 237 | 67.5 |
| | Intron1 | 153 | 35.9 |
| | Exon2 | 186 | 69.4 |
| | Intron2 | 104 | 25.0 |
| | Exon3 | 618 | 66.0 |

**Table 3 Relevant features of peroxidase sequences deduced from the gene cluster**

| | OSP1 | OSP2 | OSP3 | OSP4 | OSP5 |
|---|---|---|---|---|---|
| AA length | 338 | 335 | 336 | 348 | 346 |
| Potential signal peptide cleavage site | 28 | 28 | 31 | 24 | 22 |
| Estimated pI for mature protein | 5.76 | 8.17 | 8.18 | 4.69 | 5.28 |
| Estimated molecular mass of mature protein(kD) | 33.3 | 32.7 | 32.1 | 34.7 | 34.8 |
| Probable cellular location | Extracellular peroxidase | | | | |

的结构特征:具有一些不同于其他亚家族的保守残基,例如 Ile(I)[39]、Phe(F)[40]、Pro(P)[46]、Gly(G)[58]、Leu(L)[77]、Leu(L)[79]、Val(V)[87]、His(H)[88]等,以及一个比较保守的特征性 C 末端序列:A(T)AS(D)A(M/P)。这表明,OSP1～5、AP1、PRX7 形成了第三类植物过氧化物酶家族中一个新的分枝。

### 2.6 过氧化物酶的比较分析

对 5 个过氧化物酶及家族中部分有代表性的成员进行了两两比较分析。五个过氧化物酶间的相似程度按 OSP1～OSP5 呈梯度下降(表 4):*osp*1 与其下游 4 个基因在氨基酸水平的相同率依次为:87 %、75 %、61 %、53 %;*osp*2 与其下游基因的相同率为:72 %、56 %、51 %;OSP3 与 OSP4、OSP5 的相同率分别为 56 %、51 %。该结果表明,一系列的串联基因复制导致了该基因簇的形成,即在进化过程中依次发生了 *osp*5-*osp*4-*osp*3-*osp*2-*osp*1 的基因复制事件。

将 5 个过氧化物酶氨基酸序列用于蛋白质数据库检索发现,OSP5 与来自玉米的阴离子过氧化物酶 AP1(GenBank 登录号 Y13905)[22]相同率为 72.54 %,与来自大麦(*Hordeum vulgare*)的 PRX7(GenBank 登录号 AJ003141)[23]相同率为 63.02 %,而 OSP1、OSP2、OSP3、OSP4 与数据库中所有过氧化物酶的相同率在 55 %以下。这表明 *osp*5、*ap*1、*prx*7 为潜在的直向同源基因。

对 OSP1～5,AP1,PRX7,以及第三类过氧化物酶中几个亚家族代表进行进化分析(图 4),结果显示,*osp*1～5 在进化上发生的先后顺序为从 *osp*5 依次到 *osp*1;而且,*osp*1～5、*ap*1、*prx*7 与其他家族代表有较大差距,可归为一个新的分枝。

## 3 讨论(Discussion)

到目前为止,所发现的第三类过氧化物酶基因都具有 1～4 个外显子,而且内含子插入位点严格保守。在日本构建的水稻高密度遗传图谱中共鉴定

**Table 4 Percentage similarities and identities between the clustered peroxidases and representatives from the class III peroxidase family**

| | OSP1 | OSP2 | OSP3 | OSP4 | OSP5 | PRX7 | AP1 | BP1 | WP1 | TOPA | HRPC1 | TAP1 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 87.13 | 75.38 | 60.84 | 53.75 | 49.39 | 55.09 | 47.61 | 43.79 | 38.11 | 38.65 | 33.12 |
| OSP1 | | 89.22 | 78.42 | 65.66 | 58.26 | 54.55 | 59.28 | 52.70 | 51.63 | 45.93 | 46.32 | 38.80 |
| | | | 71.73 | 56.97 | 51.66 | 47.85 | 54.46 | 50.15 | 43.61 | 37.05 | 35.15 | 34.18 |
| OSP2 | | | 76.29 | 63.03 | 57.40 | 53.07 | 58.77 | 54.10 | 50.82 | 44.59 | 42.42 | 40.51 |
| | | | | 56.46 | 51.52 | 46.11 | 55.79 | 46.34 | 40.26 | 36.63 | 38.34 | 33.97 |
| OSP3 | | | | 62.16 | 57.62 | 50.78 | 59.76 | 52.44 | 48.19 | 44.22 | 46.32 | 42.22 |
| | | | | | 55.82 | 48.77 | 54.73 | 48.34 | 43.36 | 35.29 | 36.96 | 36.45 |
| OSP4 | | | | | 60.00 | 53.37 | 60.06 | 52.57 | 50.00 | 43.79 | 44.41 | 43.98 |
| | | | | | | 63.02 | 72.54 | 48.83 | 40.51 | 38.66 | 38.02 | 32.41 |
| OSP5 | | | | | | 67.46 | 76.59 | 53.51 | 45.34 | 46.01 | 43.71 | 38.89 |
| | | | | | | | 64.41 | 45.73 | 43.41 | 38.99 | 38.21 | 32.69 |
| PRX7 | | | | | | | 67.65 | 50.61 | 47.27 | 46.23 | 43.58 | 37.82 |
| | | | | | | | | 55.09 | 46.05 | 38.80 | 41.42 | 33.44 |
| AP1 | | | | | | | | 59.28 | 51.97 | 45.74 | 46.45 | 38.70 |
| | | | | | | | | | 45.16 | 43.18 | 39.00 | 34.04 |
| BP1 | | | | | | | | | 50.97 | 51.11 | 46.33 | 41.64 |
| | | | | | | | | | | 46.45 | 50.64 | 42.05 |
| WP1 | | | | | | | | | | 53.87 | 55.77 | 49.34 |
| | | | | | | | | | | | 49.07 | 42.62 |
| TOPA | | | | | | | | | | | 57.45 | 48.53 |
| | | | | | | | | | | | | 38.75 |
| HRPC1 | | | | | | | | | | | | 44.38 |

Percent amino acid identities are shown in row 1 and similarities are shown in row 2. Representatives from class III peroxidase family were included: BP1 (M73234)[24], WP1 (X56011)[25], TOPA (J02979)[26], HRPC1 (M37156)[27], TAP1 (X15853)[28].

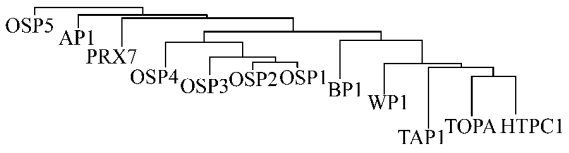


**Fig. 4 Phylogenetic analysis of amino acid sequences of OSP1-5 and other class III peroxidase family members**

The tree was constructed using GrowTree (GCG package 10.1).

出 26 个过氧化物酶位点[11]。本文对其中的一个位点标记 (R2184S) 所锚标的 BAC 克隆 q3037 (H0207F01) 进行了全序列分析，发现除 R2184S 所对应的过氧化物酶基因以外，在其上游还存在 4 个紧密排列的过氧化物酶基因，5 个基因具有相同的转录方向 (图 1)。5 个基因分别命名为 *osp*1、*osp*2、*osp*3、*osp*4、*osp*5，其中 *osp*5 对应于位点标记 R2184S。这是到目前为止所报道的最大的一个过氧化物酶基因簇。

在大规模基因组测序过程中，利用数据库中大量积累的 EST 数据进行基因组序列中基因结构的注解已变得越来越重要[19]。本文在软件预测的基础上，再结合与相应 EST 的匹配，准确地鉴定了 5 个过氧化物酶基因的外显子/内含子组织结构：*osp*1、*osp*2、*osp*3 分别含 1 个内含子，而 *osp*4、*osp*5 分别含 2 个内含子。所有内含子插入位点与其他第三类过氧化物酶基因一致[3,10]，而且内含子都具有较高的 AT 含量。高 AT 含量是植物基因内含子所普遍具有的一个特征，被认为与内含子的有效剪切有关[20]。5 个基因中多数内含子所遵循的是 GC/AG 的剪切模式，这种模式在较少一些基因中被观察到过[4]。

从与 *osp*1、*osp*3、*osp*5 相匹配的 EST 来源，还可揭示 *osp*1、*osp*3、*osp*5 的组织表达差异 (表 1)。*osp*1、*osp*3 可在根、叶等组织表达，而 *osp*5 主要在根中表达。而数据库中却没有与 *osp*2 和 *osp*4 相匹配的 EST，这表明 *osp*2 和 *osp*4 可能不表达或者极低表达，也可能在极为特殊的条件下表达，这有待于进一步的研究确证。

5 个基因分别编码 338、335、336、348 和 346 氨基酸残基的蛋白质，都具有 N 端信号肽序列，属于分泌性植物过氧化物酶。5 个蛋白质中 OSP2、OSP3 为阳离子过氧化物酶，而 OSP1、OSP4、OSP5 为阴离子过氧化物酶。但在数据库检索时我们发现与 *osp*5 相匹配的 *oscpx*1 (GenBank 登录号 AF019743) 被注解为阳离子过氧化物酶基因。经仔细比较后发现，在相对于 *osp*5 终止密码上游 41 位处 *oscpx*1 多

引入了一个 t,由此导致了移框,使得表达产物成为阳离子过氧化物酶。根据几方面的证据判断,我们倾向认为 *oscpx*1 的序列存在错误: *osp*5 的序列来源于高丰度、高质量的基因组测序,而 EST 或 cDNA 序列通常为低丰度的、单向的测序结果;数据库中相应的 EST 在该位点上与 *osp*5 一致[图 2(E)]。

5 个过氧化物酶都具有植物过氧化物酶的一些基本结构特征,如两个保守的血红素结合区(B 和 F 区,图 3),以及功能未知的 D 区;同时,又具有一些新的结构特征,如一些其他过氧化物酶所不具有的保守残基,以及一个保守的特征性的 C 末端序列(图 3)。再结合对过氧化物酶氨基酸序列的两两比较分析和进化分析结果,我们认为:本文所报道的过氧化物酶基因簇起源于 *osp*5,经一系列的串联基因复制事件所形成,同时在复制过程中丢失了一个内含子(*osp*4 ~ *osp*3),在过氧化氢酶基因家族(*cat*)进化过程中也曾发生过类似的现象[29];*osp*5 与 *ap*1、*prx*7 为潜在的直向同源基因,分别在各自的基因组中行使相同的功能;*osp*1 ~ 5、*ap*1 和 *prx*7 构成了第三类过氧化物酶基因家族中一个新的分枝。

另外,我们观察到:OSP5 与 AP1 的相似性(76.59 %)高于 OSP5 与 OSP4 的相似性(60.00 %)、OSP4 与 OSP3(62.16 %)的相似性;而与 OSP3 和 OSP2 的相似性水平相近。由此我们推测,*osp*5-*osp*4-*osp*3 的基因复制事件发生在水稻和玉米基因组分化之前,即在现在的玉米基因组中应当存在一个与本文报道的基因簇相对应的基因簇,在该基因簇中除 *ap*1 以外,至少还存在分别与 *osp*4、*osp*3 互为直向同源基因的两个过氧化物酶基因。同理可推测,在大麦基因组中也存在一个相应的基因簇,其中除 *prx*7 以外,至少还存在与 *osp*4 互为直向同源基因的过氧化物酶基因。以上推论建立在以下的前提上:不同的基因在不同的基因组中进化速率大致恒定。至于真实情况如何则有待于进一步研究确定。

植物过氧化物酶家族非常庞大,阐明整个家族的结构及进化有赖于更多家族成员的发现,而对结构进化的了解将会对认识基因功能及演化提供线索。尽管随着大规模基因组测序的展开,数据库中积累了大量的过氧化物酶基因数据,但距离全局性的认识过氧化物酶基因家族还有一定差距。本文在现有的基础上,通过对水稻第 4 号染色体长臂近端粒区一个过氧化物酶基因簇的结构及进化分析,为认识过氧化物酶基因家族的结构及进化提供了一种范例。

## References


1 Dunford H B. Horseradish peroxidase: Structure and kinetic properties. Everse J, Everse K E, Grisham M B eds. *Peroxidases in Chemistry and Biology*, Boca Raton: CRC press, 1991, **2**: 1—23

2 Welinder K G. Superfamily of plant, fungal and bacterial peroxidase. *Curr Opin Struct Biol*, 1992, **2**: 388—393

3 Justesen A F, Jespersen H M, Welinder K G. Analysis of two incompletely spliced *Arabidopsis* cDNAs encoding novel types of peroxidase. *Biochi et Biophy Acta*, 1998, **1443**: 149—154

4 Sutliff T D, Huang N, Litts J C, Rodriguez R L. Characterization of an α-amylase multigene cluster in rice. *Plant Molec Biol*, 1991, **16**: 579—591

5 Dean C, Pichersky E, Dunsmuir P. Structure, evolution, and regulation of *rbcs* genes in higher plants. *Annu Rev Plant Physiol Plant Molec Biol*, 1989, **40**: 415—439

6 Hillebrand H, Bartling D, Weiler E W. Structural analysis of the *nit*2/*nit*1/*nit*3 gene cluster encoding nitrilases, enzymes catalyzing the terminal activation step in indole-acetic acid biosynthesis in *Arabidopsis thaliana*. *Plant Molec Biol*, 1998, **36**: 89—99

7 He Z H, Cheeseman I, He D, Kohorn B D. A cluster of five cell wall-associated receptor kinase genes, *wak*1-5, are expressed in specific organs of *Arabidopsis*. *Plant Molec Biol*, 1999, **39**: 1189—1196

8 Maeda N, Smithies O. The evolution of multigene families: Human haptoglobin genes. *Ann Rev Genet*, 1986, **20**: 81—108

9 Johansson T, Nyman P O. A cluster of genes encoding major isozymes of ligin peroxidase and manganese peroxidase from the white-rot fungus *Trametes versicolor*. *Gene*, 1996, **170**: 31—38

10 Osakabe K, Koyama H, Kawai S, Katayama Y, Morohoshi N. Molecular cloning of two tandemly arranged peroxidase genes from *Populus kitakamiensis* and their differential regulation in the stem. *Plant Molec Biol*, 1995, **28**: 677—689

11 Harushima Y, Yano M, Shomura A, Sato M, Shimano T, Kuboki Y, Yamamoto T *et al*. A high-density rice genetic map with 2275 markers using a single $F_2$ population. *Genetics*, 1998, **148**: 479—494

12 Hong G, Qian Y, Yu S, Hu X, Zhu J, Tao W, Li W *et al*. A 120 kilobase resolution contig map of the rice genome. *DNA Seq*, 1997, **7**(6): 319—335

13 Roach J C, Boysen C, Wang K, Hood L. Pairwise end sequencing: A unified approach to genomic mapping and sequencing. *Genomics*, 1995, **26**(2): 345—353

14 Ewing B, Hillier L, Wendl M C, Green P. Base-calling of automated sequencer traces using Phred. I. Accuracy assessment. *Genome Res*, 1998, **8**(3): 175—185

15 Ewing B, Green P. Base-calling of automated sequencer traces using Phred. II. Error probabilities. *Genome Res*, 1998, **8**(3): 186—194

16 Altschul S F, Madden T L, Schaffer A A, Zhang J, Zhang Z, Miller W, Lipman D J. Gapped BLAST and PSI-BLAST: A new

generation of protein database search programs. *Nucleic Acids Res, 1997*, **25**(17): 3389—3402

17 Fickett J W. Recognition of protein coding regions in DNA sequences. *Nucleic Acids Res*, 1982, **10**(17): 5303—5318

18 Burge C, Karlin S. Prediction of complete gene structures in human genomic DNA. *J Mol Biol*, 1997, **268**: 78—94

19 Bailey L C, Searls D B, Overton G C. Analysis of EST-driven annotation in human genomic sequence. *Genome Res*, 1998, **8**: 362—376

20 Luehrsen K R, Walbot V. Addition of A- and U-rich sequence increases the splicing efficiency of a deleted form of a maize intron. *Plant Molec Biol*, 1994, **24**(3): 449—463

21 Welinder K G. Plant peroxidases. Their primary, secondary and tertiary structures, and relation to cytochrome peroxidase. *Eur J Biochem*, 1985, **151**: 497 - 504

22 Teichmann T, Guan C, Kristoffersen P, Muster G, Tietz O, Palme K. Cloning and biochemical characterization of an anionic peroxidase from *Zea mays*. *Eur J Biochem*, 1997, **247**: 826—832

23 Kristensen B K, Bloch H, Rasmussen S K. Barley coleoptile peroxidases. Purification, molecular cloning, and induction by pathogens. *Plant Physiol*, 1999, **120**(2): 501—512

24 Johansson A, Rasmussen S K, Harthill J E, Welinder K G. cDNA, amino acid and carbohydrate sequence of barley seed-spcific peroxidase BP1. *Plant Molec Biol*, 1992, **18**: 1151—1161

25 Rebmann G, Hertig C, Bull J, Mauch F, Dudler R. Cloning and sequencing of cDNAs encoding a pathogen-induced putative peroxidase of wheat (*Triticum aestivum* L). *Plant Molec Biol*, 1991, **16**(2): 329—331

26 Lagrimini L M, Burkhart W, Moyer M, Rothstein S. Molecular cloning of complementary DNA encoding the lignin-forming peroxidase from tobacco: Molecular analysis and tissue-specific expression. *Proc Natl Acad Sci USA*, 1987, **84**: 7542—7546

27 Fujiyama K, Takemura H, Shibayama S, Kobayashi K, Choi J K, Shinmyo A, Takano M *et al*. Structure of the horseradish peroxidase isozyme *c* genes. *Eur J Biochem*, 1988, **173**: 681—687

28 Roberts E, Kolattukudy P E. Molecular cloning, nucleotide sequence, and abscisic acid induction of a suberization-associated highly anionic peroxidase. *Mol Gen Genet*, 1989, **217**(2—3): 223—232

29 Frugoli J A, McPeek M A, Thomas T L, McClung C R. Intron loss and gain during evolution of the catalase family in angiosperms. *Genetics*, 1998, **149**: 355—365

# Structural Analysis of a Gene Cluster Encoding Two Cationic and Three Anionic Peroxidases from Rice Chromosome 4

CHEN Ze-Hua, ZHOU Bo, HAN Bin, QIAN Yue-Min, HONG Guo-Fan *

*(National Center for Gene Research, Chinese Academy of Sciences, Shanghai* 200233, *China)*

**Abstract** Sequence analysis of a rice BAC q3037(H0207F01) identified a cluster of five tandemly arranged peroxidase genes, *osp*1, *osp*2, *osp*3, *osp*4 and *osp*5, within a 22.5 kb region. *osp*4, *osp*5 each have three exons interrupted by two introns, while *osp*1, *osp*2 and *osp*3 each have two exons interrupted by a single intron. The five genes were predicted products of 338, 335, 336, 343 and 346 amino acid residues, respectively, including putative signal peptide sequence at the amino-termini. And OSP1, OSP4 and OSP5 were predicted to be anionic peroxidase, OSP2 and OSP3 are cationic. Comparative analysis and evolutionary analysis of the clustered genes and other peroxidase family members revealed that the gene cluster occurred by tandemly gene duplications (from *osp*5 to *osp*1); and that *osp*5, *ap*1 and *prx*7 were potential orthologies, and *osp*1-5, *ap*1 and *prx*7 constituted a novel evolutionary branch of class III peroxidases.

**Key words** peroxidase; gene cluster; tandemly gene duplication; ortholog

Received: November 8, 2000 Accepted: December 21, 2000

* Corresponding author: Tel, 86-21-64516371; Fax, 86-21-64825775; e-mail, gfhong@newnetra.ncgr.ac.cn