基因组学与应用生物学，2011 年，第 30 卷，第 6 期，第 649-656 页
Genomics and Applied Biology, 2011, Vol.30, No.6, 649-656

研究论文
An Article

HMMER 及同源比对预测大豆病程相关蛋白

王晶 [1,2*] 张丽伟 [1,3*] 刘春燕 [1] 李玉花 [2] 陈庆山 [1,3**] 胡国华 [1,3**]
1 黑龙江省农垦科研育种中心，哈尔滨，150090；2 东北林业大学生命学院，哈尔滨，150040；3 东北农业大学农学院，哈尔滨，150030
* 同等贡献作者
** 通讯作者，qshchen126@.com; hugh757@vip.163.com

摘 要 病程相关蛋白(pathogenesis related proteins, PRs)是病理或病理相关环境下诱导产生的一类蛋白，它的产生与积累是植物体应答生物或非生物胁迫的主要特征之一。近年来大量 PR 蛋白被鉴定，根据它们的结构特征，生物功能以及进化关系等将 PR 蛋白分为 14 个家族。然而，在重要的粮食和油料作物的大豆中发现的 PR 蛋白却很少，本文通过搜索拟南芥、水稻、玉米以及豆科植物所有的已有的 PR 蛋白，根据其保守结构域利用 BLAST 程序和 HMMER 程序同时预测大豆中可能存在的 PR 蛋白，通过两种方法的预测和比较整合，共得到大豆 9 个家族的 36 个 PR 蛋白序列。并对它们的连锁群分布、基因结构、基因长度及进化关系进行了详细的分析。发现 PR 家族成簇分布于 Gm05、Gm10、Gm13、Gm15、Gm17、Gm19 和 Gm20 等几个连锁群，基因普遍存在序列较短，大部分都小于 1 000 bp，且内含子数目较少，结构相对简单的特点。在 PR4 家族中，其家族成员亲缘关系都非常相近，而 PR1-4 和 PR1-3 等与该家族其它成员亲缘关系较远的情况。本研究结果预测的 PR 蛋白为大豆抗病育种以及抗病基因工程研究提供了良好的基础，同时为大豆中其它家族基因预测研究以及其它物种基因家族研究提供参考方法。

关键词 大豆, 病程相关蛋白(PRs), BLAST, 隐马尔可夫模型应用程序包(HMMER)

Prediction of Pathogenesis Related Protein in Soybean Using HMMER and BLAST

Wang Jing [1,2*] Zhang Liwei [1,3*] Liu Chunyan [1] Li Yuhua [2] Chen Qingshan [1,3**] Hu Guohua [1,3**]
1 The Crop Research and Breeding Center of Heilongjiang Land-Reclamation, Harbin, 150090; 2 College of Life Sciences, Northeast Forestry University, Harbin, 150040; 3 College of Agriculture, Northeast Agricultural University, Harbin, 150030
* The authors who contribute equally to this work
** Corresponding authors, qshchen126@.com; hugh757@vip.163.com
DOI: 10.3969/gab.030.000649

Abstract Pathogenesis related proteins (PRs) is a class of proteins which are induced in pathological or pathological conditions. The production and accumulation of PR protein in plant are the main characteristics in the responses of biotic and abiotic stress. In recent years a large number of PR proteins have been identified, which were divided into 14 functional families based on their structure, phylogenetic and biological activities. However, little PR protein has been found in soybean and cereal grain crops. In this paper we acquired 36 PR protein members of 9 families predicted through the BLAST and HMMER program with the queries for all the PR proteins in Arabidopsis, rice, corn and legumes. A comprehensive analysis has been carried out by the aspects of the *PR* gene distribution, gene structure, length, number of exons, and evolutionary relationships. The PR family clusters distributed in Gm05, Gm10, Gm13, Gm15, Gm17, Gm19 and Gm20, and several linkage groups, most structural features of gene are relatively simple, such as most sequences are shorter, less than 1 000 bp, and introns are less too. In the PR4 family, its members are very similar, and PR1-4 and PR1-3 with other members of the family are long distance. The predicted PR proteins in this paper might provide a good foundation for disease resistance in soybean

基金项目：本研究由国家自然科学基金项目(30971809)资助

breeding program and disease resistance genetic engineering, as well as provide a powerful gene prediction approach for other gene family in soybean genetics research.

Keywords Soybean (*Glycine max* L.), Pathogenesis related proteins (PRs), BLAST, HMMER

病程相关蛋白(pathogenesis related proteins, PRs)是存在于许多种植物中受病原菌侵染或一些特定化合物处理后新产生的一种或几种蛋白质，后来发现这些蛋白质都与病原菌侵染有关，称为病程相关蛋白。它们可以通过在侵染部位大量产生，形成抵御病原菌的保护屏障，来降低植物的敏感性，形成抗真菌或细菌的活性蛋白(Edreva, 2005)。病程相关蛋白在健康的植物中也有发现，根、衰老的叶子和植物开花期间都发现有病程相关蛋白的表达。功能主要包括：攻击病原物、降解细胞壁大分子释放二级(内源)激发子、分解毒素、结合或抑制病毒外壳蛋白等。最早是在烟草花叶病毒(tobacco mosaic virus, TMV)侵染烟草叶片时检测到 PR 蛋白的，起初被称为 b 蛋白，后来人们将其命名为病程相关蛋白(van Loon and van Kammen, 1970)。同一家族 PR 蛋白同源性较高且功能相近，不同家族的 PR 蛋白功能不同，大多为酶类，如几丁质酶等(温韵洁等, 2008)。PR 蛋白最初分为五大组(PR-1~PR-5)，是在烟草中通过分子遗传技术研究来分类，按照电泳迁移率来排序的。每组的成员都有相似的组成(Bol et al., 1990)。PR-1 组最丰富，达到叶片总蛋白的 1%~2%。PR-5 组为类甜蛋白(thaumatin-like protein, TLP)。可降解真菌细胞膜，对真菌，尤其是卵菌纲有很强的抵抗能力(Batalia et al., 1996)。可激活对丝氨酸肽链内切酶有抗性的蛋白质的活性。

根据 PR 蛋白的结构特点，可以将其分为 14 个家族(表 1) (van Loon et al., 1994; van Loon and van

表 1 PR 蛋白识别组成及家族分类(van Loon and van Strien, 1999)

Table 1 Recognized and proposed families of pathogenesis-related proteins (van Loon and van Strien, 1999)

PR 家族 PR family	家族成员 Type member	特性 Properties
PR-1	烟草 PR-1a Tobacco PR-1a	未知 Unknown
PR-2	烟草 PR-2 Tobacco PR-2	β-1,3-葡聚糖酶 β-1,3-glucanase
PR-3	烟草 P, Q Tobacco P, Q	几丁质酶型 Ⅰ, Ⅱ, Ⅳ, Ⅴ, Ⅵ, Ⅶ Chitinase type Ⅰ, Ⅱ, Ⅳ, Ⅴ, Ⅵ, Ⅶ
PR-4	烟草“R” Tobacco “R”	几丁质酶型 Ⅰ, Ⅱ Chitinase type Ⅰ, Ⅱ
PR-5	烟草 S Tobacco S	类甜蛋白 Thaumatin-like
PR-6	番茄抑制剂 Ⅰ Tomato inhibitor Ⅰ	蛋白酶抑制剂 Proteinase-inhibitor
PR-7	番茄 P6g Tomato P6g	胞内蛋白酶 Endoproteinase
PR-8	黄瓜几丁质酶 Cucumber chitinase	几丁质酶型 Ⅲ Chitinase type Ⅲ
PR-9	烟草“木质素形成过氧化物酶” Tobacco “lignin-forming peroxidase”	过氧化物酶 Peroxidase
PR-10	香菜“PR1” Parsley “PR1”	类核糖核酸酶 “Ribonuclease-like”
PR-11	烟草几丁质酶型Ⅴ Tobacco class Ⅴ chitinase	几丁质酶型 Ⅰ Chitinase type Ⅰ
PR-12	萝卜 Rs-AFP3 Radish Rs-AFP3	防御素 Defensin
PR-13	萝卜 Rs-AFP3 Radish Rs-AFP3	硫堇 Thionin
PR-14	大麦脂转移蛋白 Barley LTP4	脂转移蛋白 Lipid-transfer protein

Strien, 1999)。

然而，随着研究的进一步深入与完善，又将 PR 蛋白分为 17 个家族(王钧, 1995, 植物生理学通讯, 31(4): 312-317, 320)，其中 PR-15 和 PR-16 为萌发或萌发类似蛋白。目前分别在辣椒(李惠霞等, 2006)、马铃薯(田振东等, 2003)、豌豆(刘红霞等, 2010)、白毛杨(雷杨等, 2011, http://www.paper.edu.cn/index.php/default/releasepaper/content/201102-239)和小麦(张岗等, 2009)中均有大量的 PR 蛋白基因的研究，而在大豆中 PR 蛋白研究较少，仅 5 个，而有文献报道的研究仅有关于 PR1 和类甜蛋白两个(Graham, 2005)。因此大量地开发大豆中存在的 PR 蛋白为大豆抗病功能深入研究和大豆抗病育种都有非常重要的意义。

HMMER 是可以用来搜索使用统计模型或概要文件“隐马尔可夫模型”(HMM)的基因序列数据库的一个应用程序包。HMMER3 可以从 http://hmmer.wustl.edu/ 下载 HMMER 应用程序包。如果尚不知道可信的比对，则可以训练 HMM 来识别不一致的基因序列中的模式，并将它们大量应用于整个基因组或“表达序列标记”(EST)分析(Finn et al., 2011)。BLAST 程序是通过比对未知序列与数据库中的短序列来发现最佳匹配序列的。最初进行扫描就是确定匹配片段，序列的匹配程序由短序列的联配得分总和来决定。打分高的序列被认为是高度的同源的序列。

从功能基因研究的角度来讲，相关的搜索，比如从序列数据库中，找同源的序列，或者对一个个新的基因功能进行鉴定，使用 HMMER 比使用 BLAST 有着更高的灵敏度以及更高的搜索速度，但由于二者比对原理的不同结果也不尽相同。

本研究通过收集不同物种的 PR 蛋白序列，利用 HMMER 和 BLAST 的方法预测了大豆中可能存在的 PR 蛋白序列。并对他们的连锁群分布、基因结构、基因长度和进化关系进行了详细的分析。

1 结果与分析

1.1 同源比对获得候选 PR 蛋白序列

通过同源比对的方法我们获得了许多 PR 蛋白家族同源的序列，如在 PR-1 中我们共获得了 79 条，对这些 PR 蛋白进行多序列连配，寻找到具有典型同源的蛋白序列。如图 1 所示为部分 PR-1 蛋白序列联配后得到的保守结构域部分。

利用同源比对的方法我们可以获得较详细的基因的信息。表 2 为通过同源比对预测得到的 PR 蛋

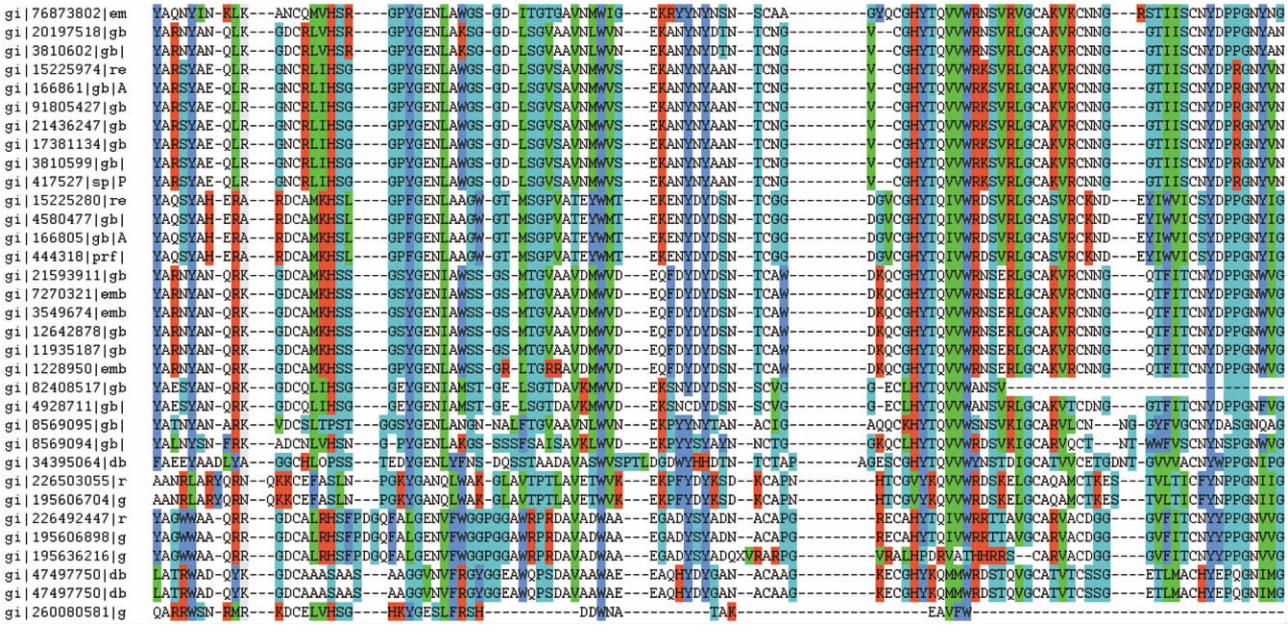

图 1 部分 PR-1 蛋白序列联配

注：**豌豆**：gi|76873802; *Glycine max*: gi|82408517, gi|4928711; **水稻**：gi|47497165, gi|47497750, gi|34395064; **玉米**：gi|260080581, gi|195636216, gi|195606898, gi|226492447, gi|195606704, gi|226503055; **其余全为拟南芥序列**

Figure 1 Sequence aligment of partial PR-1 protein

Note: *Pisum sativum*: gi|76873802; *Glycine max*: gi|82408517, gi|4928711; *Oryza sativa*: gi|47497165, gi|47497750, gi|34395064; *Zea mays*: gi|260080581, gi|195636216, gi|195606898, gi|226492447, gi|195606704, gi|226503055; Others were all *Arabidopsis thaliana* sequences

表 2 同源比对获得大豆候选 PR 蛋白序列及登陆号

Table 2 Sequence and accession number of candidate PR protein

PR 家族 PR family	数目 Number	GenBank 登录号 GenBank accession No.
PR-1	6	GR829030.1, CO985321.1, EV267541.1, EV280245.1, CK605693.1, EH222829.1
PR-2	2	CA851287.1, CK605693.1
PR-3	3	GR854577.1, EH039766.1, GR848734.1
PR-4	2	EV269439.1, GR852796.1
PR-5	6	BW655245.1, CA819895.1, EH224066.1, EV279491.1, EV280480.1, CX710653.1
PR-6	2	GR844225.1, GR844226.1
PR-10	4	FG991039.1, BI787890.1, BW653136.1, FK020996.1
PR-12	1	GR851812.1
PRNF	5	FK019450.1, GR837433.1, BW651463.1, CX703056.1, GR848614.1

白，其中 PR-1 和 PR-5 家族都预测到了 6 个同源的 PR 蛋白，而原数据库中 PR-13 和 PR-14 家族由于序列较少或同源匹配打分较低而被舍去，所以没有新的成员被预测到。

1.2 HMMER 预测获得候选 PR 蛋白的 CDS 序列

由于是对基因组的 CDS 进行预测多肽序列构建的蛋白数据库，因此利用 HMMER 方法在该数据库中预测获得候选 PR 蛋白仅仅能得到相应的多肽序列、CDS 序列及其连锁群分布和 CDS 长度(表 3)。其中 PR5 家族中预测到的成员较多，有 16 个，PR-12、PR-13 和 PR-14 家族成员有 1 个，且在这些家族中的成员普遍序列较短，大部分都小于 1 000 bp，只有 PR-12 家族中预测到了 2 589 bp 的序列。

1.3 大豆 PR 蛋白序列预测及分析

将 BLAST 和 HMMER 两种方法预测得到的序列进行比对去掉其中重复的序列，并将剩余序列进行拼接延伸，与 NCBI 进行预测注释，得到确定的 CDS。其中，PR-1、PR-3、PR-5、PR-6、PR-10 及 PRNF 数量上均有减少，PR4 家族无重复，而 PR-2 和 PR-12 由于存在与其它家族的重复而被去除，可能是由于 PR-2 与 PR-2 和 PR-1、PR-3 家族有相似性，PR-12 与 PR-13 家族存在相似性，而使重复的预测结果被去除。

PR 蛋白相对分子质量较低(6~43 kD)，在低 pH<3 下稳定，对蛋白酶有较高的抗性(van Loon and van Strien, 1999)，可以在胞内和胞间较好地积累。PR 蛋白在进化上相对保守，不同植物的同类型 PR 蛋白在分子结构和氨基酸组成等方面高度相似，因此我们以 E-value 小于 e-100 的同源序列为基因拷贝，为其进行基因定位、序列拷贝数基因数分析，以及基因长度和外显子数(表 4)，发现 PR 家族基因普遍存在序列较短，大部分都小于 1 000 bp，且内含子数目较少，结构相对简单的特点。

表 3 HMMER 预测获得候选 PR 蛋白的 CDS 序列

Table 3 CDS of candidate PR protein by HMMER

PR 家族 PR family	数目 Number	连锁群 Linkage group	CDS 长度(bp) CDS length (bp)
PR-1	3	15, 07, 10	537, 498, 762
PR-2	3	5, 15, 15	33, 814, 772, 763
PR-3	6	01, 01, 10, 16, 19, 18	63, 312, 067, 145, 108, 100, 100
PR-4	2	10, 03, 19	705, 429, 597
PR-5	16	10, 10, 11, 14, 07, 04, 11, 16, 14, 12, 15, 5, 5, 5, 15, 19	675, 567, 813, 579, 738, 1 089, 657, 1 941, 645, 702, 807, 675, 1 071, 654, 699, 963
PR-6	4	20, 12, 08, 04	477, 885, 417, 456
PR-10	6	17, 9, 15, 6, 10, 5	474, 747, 465, 543, 657, 864
PR-12	1	20	2 589
PR-13	1	17	543
PR-14	1	20	330
PRNF	2	15, 13	537, 486

表 4 PR 蛋白家族成员信息

Table 4 The information of the members in PR family

PR 家族 PR family	拷贝数 Copy	定位 Location	连锁群 Linkage group	多拷贝分布 LG & copy	E 值 E-value	长度 Length	外显子数 No. of exon
PR1-1	10	4781238-4781681	Gm15	Gm15,5; Gm13,5	e-100	444	1
PR1-2	2	4775249-4775734	Gm15	Gm15,1; Gm13,1	0	486	1
PR1-3	2	928477-928118	Gm13	Gm13,1; Gm17,1	0	360	1
PR1-4	6	2229051-2229137	Gm17	Gm17,3; Gm07,3	0	477	1
PR3-1	3	3943395-3945625	Gm02	Gm02,1; Gm16,2	0	963	3
PR3-2	3	9437955-9439341	Gm11	Gm11,1; Gm12,1; Gm13,1	0	708	2
PR3-3	3	47257733-47259129	Gm19	Gm19,1; Gm10,1; Gm02,1	0	819	2
PR4-1	1	44827166-44827694	Gm20	Gm20,1	0	453	2
PR4-2	2	49117583-49118293	Gm19	Gm19,2	0	636	2
PR4-3	1	46430142-46430949	Gm03	Gm03,1	0	429	2
PR4-4	2	49117583-49118293	Gm19	Gm19,2	0	615	2
PR5-1	3	4738945-4739662	Gm12	Gm12,1; Gm11,1; Gm20,1	0	636	1
PR5-2	3	2327444-2329498	Gm12	Gm12,1; Gm01,1; Gm11,1	0	981	2
PR5-3	2	41535645-41536319	Gm05	Gm05,2	0	675	1
PR5-4	1	46780412-46781223	Gm02	Gm02,1	1.50E-63	282	2
PR5-5	2	15933457-15934851	Gm07	Gm07,1; Gm08,1	0	738	2
PR5-6	6	5625916-5626340	Gm10	Gm10,6	0	425	1
PR5-7	2	41535954-41536319	Gm05	Gm05,2	0	366	1
PR5-8	7	5624843-5626534	Gm10	Gm10,7	0	651	3
PR5-9	2	1800851-1801525	Gm11	Gm11,1; Gm10,1	0	675	1
PR5-10	4	49305647-49307309	Gm14	Gm14,1; Gm17,1; Gm19,1; Gm04,1	0	627	2
PR5-11	4	38386057-38386794	Gm05	Gm05,2; Gm08,1; Gm12,1; Gm10,1	0	738	1
PR6-1	3	43135962-43136157	Gm20	Gm20,3	2.90E-96	213	1
PR6-2	2	14262715-14267005	Gm12	Gm12,1; Gm06,1	0	885	4
PR10-1	1	12036586-12037191	Gm15	Gm15,1; Gm06,1	2.00E-153	483	2
PR10-2	4	3355972-3357768	Gm09	Gm09,2; Gm15,2	0	702	3
PR10-3	2	3324147-3325199	Gm09	Gm09,1; Gm15,1	0	477	2
PR10-4	1	39741646-39744207	Gm20	Gm20,1	1.10E-120	672	3
PR10-5	2	12001574-12001852	Gm15	Gm15,1; Gm09,1	1.50E-139	279	1
PR10-6	1	2216965-2217779	Gm17	Gm17,1	0	474	2
PR14-1	2	35867299-35872962	Gm20	Gm20,1; Gm10,1	0	1 218	8
PRNF-1	2	922675-923202	Gm13	Gm13,1; Gm17,1	0	528	1
PRNF-2	1	35995473-35996433	Gm13	Gm13,1	2.00E-63	372	3
PRNF-3	2	60632478-60634170	Gm18	Gm18,1; Gm08,1	0	1 191	2
PRNF-4	1	36771289-36772060	Gm13	Gm13,1	1.60E-107	279	2
PRNF-5	2	4775249-4775734	Gm15	Gm15,1; Gm13,1	0	486	1

对所有的 PR 蛋白家族成员及拷贝基因进行连锁群分布研究发现：PR 蛋白基因主要集中分布于 Gm05、Gm10、Gm13、Gm15、Gm17、Gm19 和 Gm20 等几个连锁群，而其它连锁群相对较少。说明 PR 蛋白基因之间存在着成簇分布的现象特别是同一家族成员之间成簇现象更为严重，如图 3 所示的大部分的 PR5 家族成员分布于 Gm10 连锁群上(图 2; 图 3)。

利用 MEGA4 对获得的这 9 个家族 36 个成员进行进化分析，同一家族成员间大部分进化起源较为相近，如 PR4 家族亲缘关系都非常相近，同时也存在相同家族进化关系较远的情况，如 PR1-4 和 PR1-3 等与该家族其它成员亲缘关系较远。而未分类家族可能由某些家族进化而来(图 4)。

2 讨论

2.1 PR 蛋白基因预测的意义和可行性

PR 蛋白基因的表达受病原菌侵染、植物发育阶

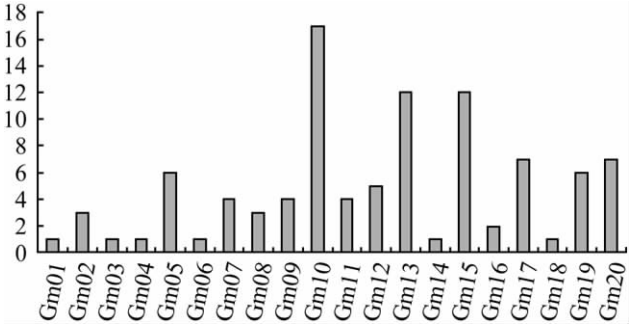

图 2 PRs 基因在连锁群上的分布

Figure 2 Distribution of genes in PRs families on the LGs

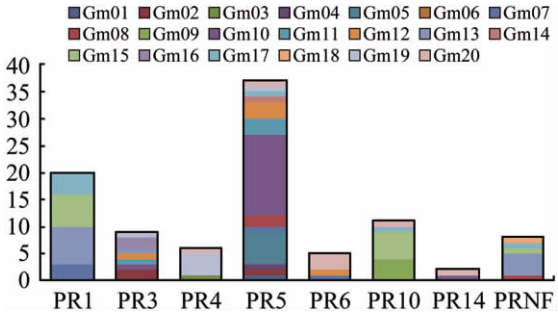

图 3 PRs 家族分布的连锁群

Figure 3 Distribution of PRs families on the LGs

段、激素和胁迫等因素的调节，参与植物的局部和系统诱导抗性。然而，目前对 PR 蛋白基因的表达调控机理及引起 PR 蛋白基因表达的信号传导途径知之甚少。所以，PR 蛋白基因的预测对 PR 蛋白基因深入研究抗病调节过程有着重要的作用。而基于 PR 蛋白序列的保守性和同一家族功能结构域的同一性，使得同样基于相同原理的同源比对和特征分析的 HMMER 成为有力的预测方法，同时通过二者结合，重复序列的发现相互验证了彼此预测的准确性。但这两种预测方法同时存在着预测结果较少的现象，可能是由于 3 种原因：(1) PR 蛋白家族成员间同源性很高，就使得我们要预测的源序列减少；(2)由于对同源性要求较高，打分低的序列就被排除在外；(3) HMMER 的预测原理是整合序列之间的共同特征，再根据这一特征来搜索序列，这也就又使得源序列减少。

2.2 PR 蛋白特征分析

PR 蛋白基因表达的基本机制是转录活化。同一家族成员间的大部分具有相同的内含子数，同一拷贝之间内含子数也完全相同。不同长度的内含子，内含子的长短与是否接受某种信号有关。大部分成员之间比较集中地分布于几个连锁群上，成簇分布，可能利于在接受刺激信号后的强烈表达或在病原菌侵染后的不断连续产生大量的 PR 蛋白。并且不同基因

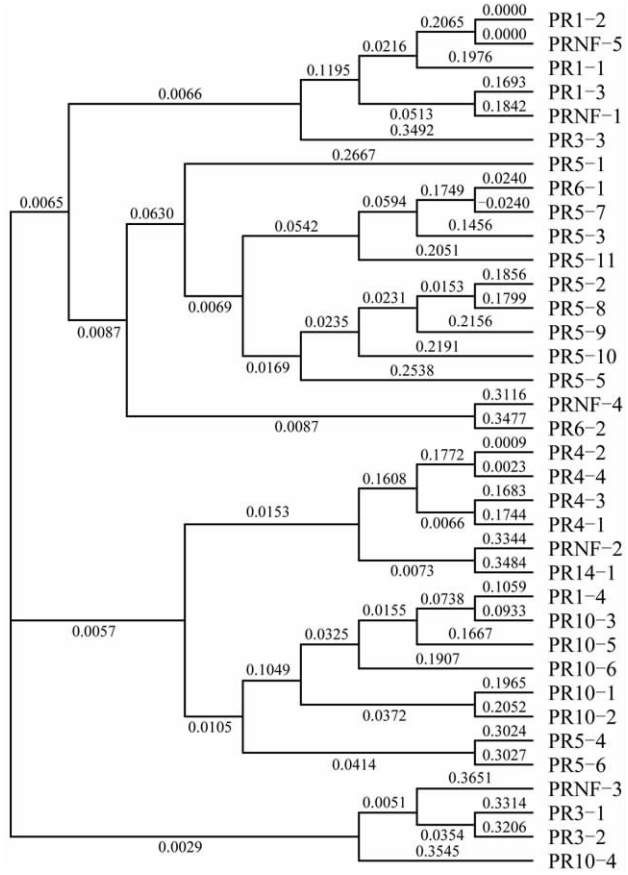

图 4 PR 蛋白基因的聚类分析

Figure 4 The cluster analysis of PR protein gene

家族可接受同一刺激信号而激活表达，可能是由于不同家族之间也存在着同源性，导致有些家族在预测中由于与其它家族预测的重复而被去除；而对同一种刺激信号，不同家族的激活表达可是同步的和协调的，也可是相互抑制的。靶位序列对不同刺激信号表现出特异性，例如番茄中有 3 种不同的 PR 蛋白对氨基丁酸的 3 种异构体反应差异达 86%，说明不同家族间的差异序列可能决定着 PR 蛋白的特异性(赵淑清和郭剑波, 2003)。

3 材料和方法

3.1 PR 蛋白序列信息收集和整理

按家族分类分别从 NCBI (http://www.ncbi.nlm.nih.gov/)上下载拟南芥(*Arabidopsis thaliana*)、玉米(*Zea mays*)、水稻(*Oryza sativa*)以及豆科植物(*Fabaceae*)的 PR 蛋白序列。

共下载拟南芥 PR 蛋白 114 个、水稻 83 个、玉米 23 个及豆科植物 46 个，其中大豆的 PR 蛋白仅有 4 个。将他们按家族名称分类，共得到 PR-1、PR-2、PR-3、PR-4、PR-5、PR-6、PR-10、PR-14 及未定

家族(pathogenesis-related protein in no family, PRNF) PR 蛋白等 9 类 266 个 PR 蛋白序列(表 5)。

3.2 大豆蛋白数据库及软件的准备

从 NCBI (http://www.ncbi.nlm.nih.gov/) 上下载大豆基因组数据库，运用 GENSCAN 对大豆基因组进行开放阅读框(ORF)的预测，并将得到全基因组的编码序列预测其蛋白序列，建立蛋白质数据库。同时下载大豆的 EST 数据库。

从 NCBI 下载用于进行本地比对的BLAST2.2.16 软件包，并下载 HMMER3.0 软件进行安装。

3.3 PR 蛋白通过同源比对预测

将下载的 PR 蛋白序列进行去重复，对于 E 值小于 0.01 的序列可以认定为同源序列，将其去除，按照与大豆亲缘关系由近及远的选择方式仅保留一条非重复序列。将去重复的 PR 蛋白序列利用 tblastn 程序与大豆 EST 数据进行比对，获得与 PR 蛋白序列同源的大豆 EST 序列，即得到候选大豆 PR 蛋白 EST 序列。

3.4 PR 蛋白的 HMMER 预测

将下载的拟南芥、玉米、水稻以及豆科植物的 PR 蛋白序列按家族分别作多序列联配，得到 ALIGN 文件，并转换成 HMMER 可识别的文件，并分别保存为 seed 和 align 文件。对于成员较少或多序列联配后同源性较差的家族，则通过 NCBI 同源搜索找出网络数据库中其它同源基因，通过对其它同源基因的多序列联配，得到 seed 文件。

通过 HMMbuild 将 align 文件和 seed 文件转换成隐马尔科夫模型文件 seed.hmm 和 align.hmm，建立 PR 蛋白各家族家族的隐马尔科夫模型。

程序命令为“# hmmbuild PR.hmm PR.msf”。

通过 HMMsearch 用已建立的 HMM 文件对先前预测的大豆蛋白数据进行比对，E 值设定为 HMMER 默认值 E-value 0.01，得出.out 文件。

程序命令为“#hmmsearch PR.hmm soybeandatabase>PR.out”。

再根据输出的 out 文件返回构建的本地蛋白数据库寻找到预测的 PR 蛋白的多肽序列和 CDS 序列，作为候选大豆 PR 蛋白。

3.5 大豆 PR 蛋白序列预测及分析

将两种方法预测的候选大豆 PR 蛋白 EST 序列进行整合，去掉重复预测的序列。再对其候选序列进行多轮的拼接电子延伸，并用 GENSCAN (http://genes.mit.edu/GENSCAN.html)预测全长 ORF。将预测的全长 ORF 与 NCBI 进行同源比对，进行基因功能注释以及确定真正的 CDS，并将其按家族和顺序进行分类，命名为大豆 PR 蛋白。

利用 Phytozome (http://www.phytozome.net/)对得到的大豆 PR 蛋白在大豆基因组上进行定位，同时确定其在基因组上的分配情况、拷贝数、外显子内含子数，以及多拷贝基因间结构变异和进化。

表 5 PR 家族分类及成员数目

Table 5 Classification and numbers of PR families

PR 蛋白家族 PR family	成员数目 Number	拟南芥 *Arabidopsis thaliana*	玉米 *Zea mays*	水稻 *Oryza sativa*	豆科植物 *Fabaceae*
PR-1	79	26	12	29	12
PR-2	4	0	0	0	4
PR-3	17	3	0	13	1
PR-4	4	1	1	1	1
PR-5	19	14	3	0	2
PR-6	3	3	0	0	0
PR-10	29	0	3	4	22
PR-12	4	4	0	0	0
PR-13	6	6	0	0	0
PR-14	14	14	0	0	0
PRNF	87	43	4	36	4
总计 Totle	266	114	23	83	46

注: PRNF 为 pathogenisis related protein in no family

Note: PRNF means pathogenisis related protein in no family

作者贡献

陈庆山老师负责实验设计和指导；王晶和张丽伟负责了软件分析、数据整理和论文写作；刘春燕、李玉花和胡国华老师帮助论文修改。

致谢

本研究由国家自然科学基金项目(30971809)资助，且得到朱命喜同学的指导和大力支持，特此致谢！

参考文献

Batalia M.A., Monzingo A.F., Roberts W., and Robertus J.D., 1996, The crystal structure of the antifungal protein zeamatin, a member of the thaumatin-like, PR-5 protein family, Nature Struct. Biol., 3(1): 19-23

Bol J.F., Linthorst H.J.M., and Cornelissen B.J.C., 1990, Plant pathogenesis—related proteins induced by virus infection, Annu. Rev. Phytopathol., 28: 113-138

Edreva A., 2005, Pathogenesis-related proteins: Research progress in the last 15 years, Gen. Appl. Plant Physiology, 31(1-2): 105-124

Finn R.D., Clements J., and Eddy S.R., 2011, HMMER web server: Interactive sequence similarity searching, Nucleic Acids Research, 39(Web Server Issue): W29-W37

Graham M.Y., 2005, The diphenylether herbicide lactofen induces cell death and expression of defense-related genes in soybean, Plant Physiol., 139(4): 1784-1794

Li H.X., Xie B.Y., and Feng L.X., 2006, Accumulation of pathogenesis-related proteins and their activities of pepper plants induced by β-aminonbutyric acid, Yuanyi Xuebao (Acta Horticulturae Sinica), 33(6): 1335-1337 (**李惠霞**, **谢丙炎**, **冯兰香**, 2006, β- **氨基丁酸诱导辣椒产生** PR **蛋白及其酶活性的变化**, **园艺学报**, 33(6): 1335-1337)

Liu H.X., Zhao X., Bi Y., Zhang Z.Y., Chen B.H., and An C.C., 2010, A proteomic approach to study defense-related proteins responses to GSH treatment in pea (*Pisum sativum*), Zhongguo Nongye Kexue (Scientia Agricultura Sinica), 43(22): 4746-4753 (**刘红霞**, **赵鑫**, **毕阳**, **张增艳**, **陈佰鸿**, **安成才**, 2010, **豌豆病程相关蛋白应答** GSH **的蛋白质组学分析**, **中国农业科学**, 43(22): 4746-4753)

Tian Z.D., Liu J., and Xie C.H., 2003, Cloning of a pathogenesis-related protein gene cDNA of potato using RACE methods combined with cDNA library, Yichuan Xuebao (Acta Genetica Sinica), 30(11): 996-1002 (**田振东**, **柳俊**, **谢从华**, 2003, cDNA **文库与** RACE **方法结合克隆一个马铃薯病程相关蛋白基因** cDNA, **遗传学报**, 30(11): 996-1002) (Chinese journal in English)

van Loon L.C., Pierpoint W.S., Boller T., and Conejero V., 1994, Recommendations for naming plant pathogenesis-related proteins, Plant Mol. Biol. Rep., 12(3): 245-264

van Loon L.C., and van Kammen A., 1970, Polyacrylamide disc electrophoresis of the soluble leaf proteins from *Nicotiana tabacum* var. Samsun and Samsun NN. Ⅱ. changes in protein constitution after infection with tobacco mosaic virus, Virology, 40(2): 190-211

van Loon L.C., and van Strien E.A., 1999, The families of pathogenesis-related proteins, their activities, and comparative analysis of PR-1 type proteins, Physiol. Mol. Plant Pathol., 55(2): 85-97

Wen Y.J., He H.W., Huang Q.S., Liang S., and Bin J.H., 2008, Roles of pathogenesis-relative protein 10 in plant defense response, Zhiwu Shenglixue Tongxun (Plant Physiology Communications), 44(3): 585-592 (**温韵洁**, **何红卫**, **黄群生**, **梁山**, **宾金华**, 2008, **病程相关蛋白** 10 **在植物防御反应中的作用**, **植物生理学通讯**, 44(3): 585-592)

Zhang G., Li Y.M., Zhang Y., Dong Y.L., Wang X.J., Wei G.R., Huang L.L., and Kang Z.S., 2009, Cloning and characterization of a pathogenesis related protein gene *TaPR10* from wheat induced by stripe rust pathogen, Zhongguo Nongye Kexue (Scientia Agricultura Sinica), 42(1): 110-116 (**张岗**, **李依民**, **张毅**, **董艳玲**, **王晓杰**, **魏国荣**, **黄丽丽**, **康振生**, 2009, **条锈菌诱导的小麦病程相关蛋白** *TaPR10* **基因的克隆及特征分析**, **中国农业科学**, 42(1): 110-116)

Zhao S.Q., and Guo J.B., 2003, Systemic acquired resistance and signal transduction in plant, Zhongguo Nongye Kexue (Science Agriculture Sinica), 36(7): 781-787 (**赵淑清**, **郭剑波**, 2003, **植物系统性获得抗性及其信号转导途径**, **中国农业科学**, 36(7): 781-787)