Computational Molecular Biology (online), 2011, Vol.1
ISSN 1927-5587 http://cmb.sophiapublisher.com

Research Article Open Access

Pathogenesis Related Protein (PR protein) in Soybean Predicted Through HMMER and BLAST Resources

Jing Wang [1,2], Liwei Zhang [1,3], Chunyan Liu [1], Yuhua Li [2], Qingshan Chen [1,3], Guohua Hu [1,3]
1. The Crop Research and Breeding Center of Heilongjiang Land-Reclamation, Harbin 150030, P.R. China;
2. College of Life Sciences, Northeast Forestry University, Harbin 150040, P.R. China;
3. College of Agriculture, Northeast Agricultural University, Harbin 150090, P.R. China
Corresponding authors, qshchen@126.com; Hugh757@vip.163.com; Authors
Computational Molecular Biology 2011, Vol.1 No.2 DOI: 10.5376/cmb.2011.01.0002
Received: 14 Oct., 2011
Accepted: 26 Nov., 2011
Published: 29 Dec., 2011

Preferred citation for this article as:
Wang et al., 2011, Pathogenesis Related Protein (PR Protein) in Soybean Predicted Through HMMER and BLAST Resources, Genomics and Applied Biology, 30(6): 649-656 (doi:10.3969/gab.030.000649)

Abstract The production and accumulation of PR (pathogenesis related protein) 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 extron, and evolutionary relationships. The predicted PR proteins in this paper might provide a good foundation for disease resistance in soybean 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

Background

There are several new proteins which were present in many species of plants infected with the pathogen or induced by some specific compounds, all of these proteins have been found infected with the pathogen later, and known as pathogenesis-related protein. They have the ability of anti-fungal or bacterial, when a large number of these proteins were produced in the infection site, forming a protective barrier against pathogens to reduce the sensitivity of plants (Edreva, 2005). PR protein was detected when the tobacco mosaic virus (TMV) infected tobacco leaves initially, firstly called the b-protein, and then renamed pathogenesis-related proteins (van Loon and van Kammen, 1970). The PR proteins of the same family had higher homology sequences and the similar function, contrariwise, they have different functions, and most of them were enzymes, such as chitinase (Wen et al., 2008). PR protein was originally divided into five groups (PR−1 to PR−5), in the research of tobacco. They were classified by molecular genetic techniques, sorted according to the electrophoretic mobility. Each member in one group has a similar composition (Bol et al., 1990). PR−1 group, the most abundant, reached 1%~2% of total leaf protein. PR−5 group was thaumatin-like protein (TLP), which could biodegradable fungal cell membrane, especially had a strong resistance to Oomycetes (Batalia et al., 1996). And it can activate the activity of the resistant protein to serine endopeptidase enzyme.

According to structural features, the PR protein can be divided into 14 families (Table 1) (van Loon et al., 1994; van Loon and van Strien, 1999).

However, as the further research and its improvement, the PR proteins were divided into 17 families (Wang, 1995), in which PR15 and PR16 were similar to germination or germin-like protein. At present, only five PR proteins in soybean were identified. While only PR1 and thaumatin-like protein were reported in the references (Graham, 2005). Therefore, it has a very important significance to predict and investigate PR protein of soybean.

Computational Molecular Biology (online), 2011, Vol.1
ISSN 1927-5587 http://cmb.sophiapublisher.com

Table 1 Recognized and proposed families of pathogenesis-related proteins

PR family	Type member	Properties
PR−1	Tobacco PR−1a	Unknown
PR−2	Tobacco PR−2	β−1,3−glucanase
PR−3	Tobacco P,Q	Chitinase type I, II, IV, V, VI, VII
PR−4	Tobacco “R”	Chitinase type I, II
PR−5	Tobacco S	Thaumatin-like
PR−6	Tomato Inhibitor I	Proteinase-inhibitor
PR−7	Tomato P6g	Endoproteinase
PR−8	Cucumber chitinase	Chitinase type III
PR−9	Tobacco “lignin-forming peroxidase”	Peroxidase
PR−10	Parsley “PR1”	“Ribonuclease-like”
PR−11	Tobacco class V chitinase	Chitinase type I
PR−12	Radish Rs-AFP3	Defensin
PR−13	Radish Rs-AFP3	Thionin
PR−14	Barley LTP4	Lipid-transfer protein

In this paper, we predicted the candidate PR protein sequence in soybean by BLAST and HMMER methods based on assembled the PR protein sequences from different species. And a detailed analysis was made on the linkage group of PR proteins and their distribution, gene structure, gene length, and the relationship between the evolutions.

1 Results and Analysis

1.1 Obtain candidate PR protein sequences by BLAST

A lot of PR proteins homologous sequences were obtained by BLAST method, such as in PR-1 family, we gained 79 PR proteins and found a protein sequence with a typical domain by multiple sequence alignments. Figure 1 showed the conserved domain of PR1 partial protein via sequence alignment.

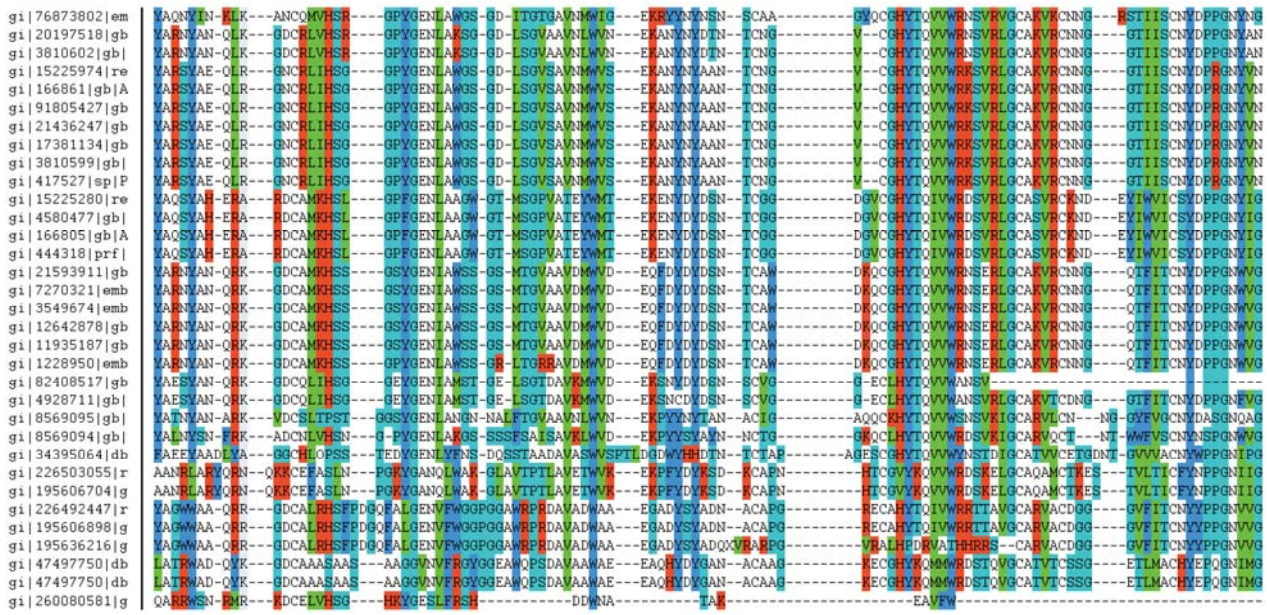

Figure 1 Sequence alignment of partial PR1 protein

More detailed genetic information was obtained by the method of BLAST. Table 2 showed the PR proteins which were obtained by BLAST, assembling, and prediction. In PR1 and PR5 family, both of them were predicted six homologous PR proteins, while in the original database, PR13 and PR14 family were discarded, because either there were not enough sequences or the lower score in homology match.

Table 2 Sequence and accession number of candidate PR protein

PR family	Number	GenBank accession No.
PR1	6	GR829030.1,CO985321.1, EV267541.1, EV280245.1, CK605693.1, EH222829.1
PR2	2	CA851287.1, CK605693.1
PR3	3	GR854577.1, EH039766.1, GR848734.1
PR4	2	EV269439.1, GR852796.1
PR5	6	BW655245.1, CA819895.1, EH224066.1,EV279491.1, EV280480.1, CX710653.1
PR6	2	GR844225.1, GR844226.1
PR10	4	FG991039.1, BI787890.1, BW653136.1, FK020996.1
PR12	1	GR851812.1
PRNF	5	FK019450.1, GR837433.1, BW651463.1, CX703056.1, GR848614.1

1.2 Obtain the CDS of candidate PR protein sequences by HMMER

Only CDS of the candidate PR protein sequences can be obtained, because the database was established with the CDS of the genome. The linkage group distribution and length of CDS can be obtained by the HMMER program (Table 3).

Table 3 CDS of candidate PR protein sequences by HMMER

PR family	Number	Gm	CDS length
PR1	3	15,07,10	537, 498, 762
PR2	3	5,15,15	33,814,772,763
PR3	6	01,01,10,16,19,18	63,312,067,145,108,100,000
PR4	2	10,03,19	705,429,597
PR5	16	10,10,11,14,07,04,11,16,14,12,15,5,5,5, 15,19	675,567,813,579,738,1089,657,1941,645,702,807,675,1071, 654,699,963
PR6	4	20,12,08,04	477,885,417,456
PR10	6	17,9,15,6,10,5	474,747,465,543,657,864
PR12	1	20	2589
PR13	1	17	543
PR14	1	20	330
PRNF	2	15,13	537,486

1.3 The prediction and analysis of soybean PR protein sequences

The sequences from the two prediction methods were alignment, and the repetitive sequences were removed. The unique sequences were assembled and extended. And then we submitted the assembling sequences to NCBI for annotation, at last, we determined the real CDS sequences. Among them, the members of PR1, PR3, PR5, PR6, PR10, and PRNF all have reduced, PR4 family has no repetitive, while PR2 and PR12 families have been removed because both the sequences of prediction were just the duplication of

other families, and these two families were not shown in the later results.

The character of PR proteins with low molecular weight (6~43 kD) was stable when the pH value was lower than 3.0, and it has a higher resistance to the protease, so the PR proteins are established in all plant organs – leaves, stems, roots, flowers, particularly abundant in the leaves. PRs have dual cellular localization – vacuolar and apoplastic, the apoplast being the main site of their accumulation (van Loon, 1999). Apart from being present in the primary and secondary cell walls of infected plants, PRs are also found in cell wall appositions (papillae) deposited at the inner side or the space of cell wall in response to fungal attack. It was relatively conserved in evolution, the same type of PR proteins of different plants were highly similar in the molecular structure, amino acid composition and so on. So the E-value of homologous sequences less than e−100 was seemed as the gene copies. Table 4 showed the gene mapping, the number of sequence copy, the analysis number and length of gene, and their number of exon.

Table 4 The information of the members in PR family

PR family	Copy	Location	LG	LG & copy	E-value	Length	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

Computational Molecular Biology (online), 2011, Vol.1
ISSN 1927-5587 http://cmb.sophiapublisher.com

Continuing Table 1

PR family	Copy	Location	LG	LG & copy	E-value	Length	No. of exon
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	1218	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	Gm18,1Gm08,1	0	1191	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

We studied all the members of PR protein family and linkage group distribution of their copies, and found that PR protein mainly distributed on Gm05, Gm10, Gm13, Gm15, Gm17, Gm19, and Gm20 linkage groups, while less on others. It demonstrated that PR protein genes were clustered distribution on the linkage groups, especially among the members in the same family, as shown in Figure 3, the most members of PR5 family located in Gm10 linkage group (Figure 2; Figure 3).

The phylogenetic analyses for thirty-six members in 9 families by using MEGA4 showed that most of the members in the same family had the similar evolutionary origin, such as the PR4 family was all clustered together and each of them was very short in distance, the distance between PR4-2 and PR4-4 was less than 0.002; moreover, there was also a far evolutionary relationships between the members in the same family, such as PR1-4 and PR1-3. However, the PRNF may have evolved from some other families (Figure 4).

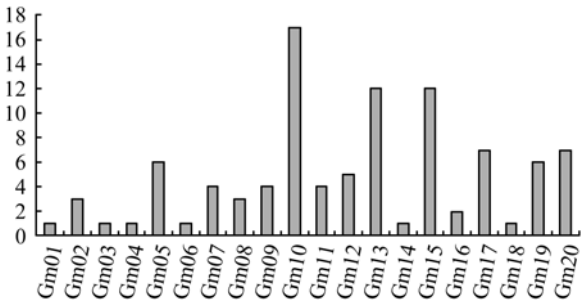

Figure 2 Distribution of genes in PRs families on the LGs

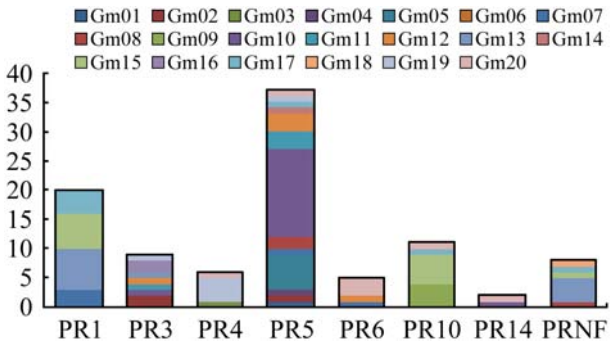

Figure 3 Distribution of PRs families on the LGs

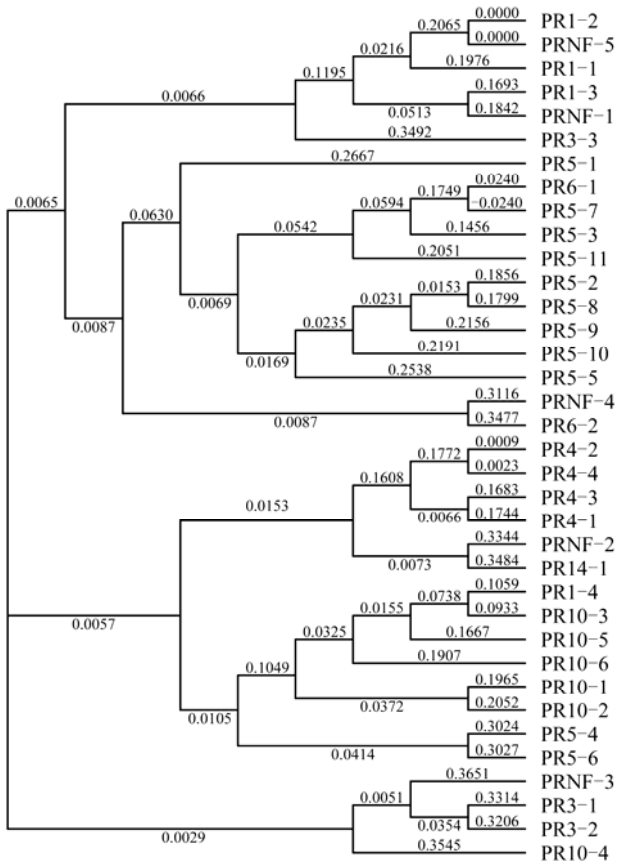

Figure 4 The cluster analysis of PR protein gene

2 Discussion

2.1The significance and feasibility of PR protein gene prediction

The expression of PR proteins was regulated by pathogen infection, plant development, and other factors such as stress or hormones, which involved in partial and systemic resistance. However, it is seldom known about the regulation mechanism of gene expression and the gene expression in the pathway of signal transduction. Therefore, HMMER has become the powerful method based on sequence alignment and characteristic analysis. Meantime, the PR protein sequences were conserved and the function of the same family was similar. The combination of HMMER and BLAST produced a lot of repeat sequences which verified the accuracy of the two methods. However, both methods present the phenomenon of less result after prediction; there might be three kinds of reasons as follows. Firstly, the higher homology among the PR protein family members, the less it makes the source for researching; secondly, because of the higher homology, the low-scoring sequence was excluded; as well as the principle of HMMER sequence prediction was also the common features integration among different sequences. And the feature was also used for sequence searching, which also reduced the source of sequence.

2.2 PR protein characterization analysis

The basic mechanism of PR protein gene expression is transcriptional activation. Most members of the same family and the different copies have the same number of intron. Different introns were different in length, which decided whether to accept some kind of signaling. Most members are clustered distribution on several linkage groups, which may induce amounts of protein production continuously after pathogen infection or may stimulate the expression after receiving a strong signal. Due to the homology in different families, they can receive the same stimulus signal to activate the expression and the repeat sequences in different families may also be deduced after prediction. The activation of expression of different families stimulated by the same signal not only synchronization and coordination, but also mutual inhibition. The target sequence performed specificity on the signal, for example, there are three kinds of tomato PR proteins, which response to the three isomers of ɣ-aminobutyric acid (GABA) differences reaching about 86%, indicating that the differences between the different families may determine the sequence specificity of PR proteins (Zhao and Guo, 2003).

3 Materials and Methods

3.1 Materials

3.1.1 The collection and collation of PR protein sequences

The Arabidopsis (*Arabidopsis thaliana*), maize (*Zea mays*), rice (*Oryza sativa*), and legumes (*Fabaceae*) PR protein sequences were downloaded from the NCBI (http://www.ncbi.nlm.nih.gov/) and classified according to their families.

In this document, the number of PR protein sequences for Arabidopsis, rice, corn, and legumes downloaded from the internet were 114, 83, 23, and 46, respectively.

Among them, only 4 PR proteins belonged to soybean. All of them were classified by their family names, and we totally obtained 266 members belonged to 9 families, which were 1, 2, 3, 4, 5, 6, 10, 14 and uncertain family (Pathogenesis-related protein in No Family, PRNF) (Table 5).

Table 5 Classification and numbers of PR families

PR family	Number	*Arabidopsis thaliana*	*Zea mays*	*Oryza sativa*	*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
Total	266	114	23	83	46

Note: PRNF means pathogenisis related protein in no family

3.1.2 Database of soybean protein and software preparation

Soybean genome database was downloaded from NCBI (http://www.ncbi.nlm.nih.gov/), the open reading frame (ORF) in the genome were predicted by GENSCAN and then were translated into protein to establish the protein database. The soybean EST database was downloaded at the same time.

Download the blast 2.2.16 package for the local alignment from NCBI, and HMMER 3.0 software for installment.

3.2 Methods

3.2.1 PR protein were predicted by BLAST

The repeat sequences of the PR protein would be removed. The sequence can be identified as the homology sequence when their E-value was less than 0.01, and only a non-repetitive sequence was retained in accordance with the soybean genetic relationship options from near to far. The unique PR protein sequence would be BLAST with the database of soybean EST by using tBLASTn procedure, and the result sequences from soybean EST that homology with the PR protein were determined to be the candidates of PR proteins.

3.2.2 HMMER predicted PR protein

The multiple alignment according to the PR family were made among the download sequences of Arabidopsis, maize, rice, and legumes to form the ALIGN file and converted it into a recognizable HMMER file. They were saved as seed and align files respectively. For the families with less members or less homologous sequences after alignment, we would search through NCBI to identify the other homology genes in the network database, and then carried out the multiple alignment to form the seed file.

The file of align and seed were transformed into hidden Markov model (HMM) file seed.hmm and align.hmm by HMMbuild, and established the HMM of the family-owned of PR proteins.

The program commanded as "# hmmbuild PR.hmm_PR.msf".

The HMM files were compared with the database of soybean protein by HMMsearch, according to the default E-value 0.01, and obtained the out file.

The program commanded as "# hmmsearch PR.hmm soybeandatabase> PR.out".

According to the out file, the predicted peptide sequences and CDS sequences of PR protein can be found in the new local protein database, as the candidate of soybean PR protein.

3.2.3 Prediction and analysis of PR protein sequence of soybean

The candidate soybean EST sequences of PR protein predicted through two ways were integrated, and the repeat sequences were removed. The candidate sequences were also assembled for several rounds and extended by GENSCAN (http://genes.mit.edu/GENSCAN.html) to predict the full-length ORF. The predicted full-length ORF were annotated by researching with the NCBI to determine the function and the real CDS. At last they were classified according to their family or order and named the soybean PR protein.

The PR protein of soybean were mapped on soybean genome by Phytozome (http://www.phytozome.net/), and determined their distribution on the genome, number of copies, the number of exon and intron, and structural and evolution variation among more copies of genes.

Author's contributions

Qingshan Chen professor was responsible for experimental design and the experiment direction; Jing Wang and Liwei Zhang were responsible for software analysis, data management and paper writing; Chunyan Liu, Yuhua Li and Guohua Hu teachers helped modified the paper.

Acknowledgements

This study was funded by the National Natural Science Foundation (30971809), and we got the help and support from Zhu Mingxi, Thanks a lot.

References

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: 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

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

van Loon L.C., and van K., 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: 199-211

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: 85-97

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

Wang J., 1995, Recent advance of plant disease resistance. Zhiwu Shenglixue Tongxun (Plant Physiology Communications), 31 (4): 312-317

Wen Y.J., Huang Q.S., Liang S., Bin J.H., and He H.W., 2008, Roles of pathogenesis-relative protein 10 in plant defense response, Zhiwu Shenglixue Tongxun (Plant Physiology Communications),44(3): 585-592

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