Characterization of a Proline-rich Cell Wall Protein Gene Family of Soybean
A COMPARATIVE ANALYSIS

Jong Chan Hong†, Ron T. Nagao, and Joe L. Key‡
From the Department of Botany, University of Georgia, Athens, Georgia 30602

Further characterization of a proline-rich cell wall protein gene family from soybean (Glycine max (L.) Merr) has been accomplished by the isolation and sequence analysis of two additional genes, SbPRP2 and SbPRP3, which encode mRNAs of 1050 and 650 nucleotides in length, respectively. Like the proline-rich protein gene, ShbPRP1, which was previously reported (Hong, J. C., Nagao, R. T., and Key, J. L. (1987) J. Biol. Chem. 262, 8367–8376), these two ShbPRP genes encode proteins having a signal peptide sequence and repeats of Pro-Pro-Val-Tyr-Lys. The SbPRP2 gene encodes a protein of 26 kDa which contains a perfect alternating repeat of Pro-Pro-Val-Tyr-Lys and Pro-Pro-Val-Glu-Lys. The SbPRP3 encodes a 10-kDa protein which also contains Pro-Pro-Val-Tyr-Lys as a major amino acid repeat, but the overall amino acid sequence of this protein is more variable than that of SbPRP1 and SbPRP2.

RNA blot analyses have demonstrated that there are marked differences in the pattern of expression of each SbPRP in various soybean tissues. In contrast, sequence analysis reveals that the SbPRP genes contain a high degree of sequence conservation. Nucleotide sequence homology extends 60 to 100 base pairs upstream of the transcription initiation site and includes typical CAT and TATA sequences. Approximately 80 base pairs of the 3′-noncoding sequence around the polyadenylation signal is also highly conserved. Therefore, the DNA sequence upstream of the 5′-conserved region is presumed to contain cis-elements accounting for the developmental and tissue specificity of gene expression. While the pentameric repeat structures occur in all SbPRP genes, the encoded proteins are predicted to be different in several features including basicity, substitutions of tyrosine and glutamic acid in the repeat, and the size of the mature protein.

Several plant cDNAs or genes that encode structural cell wall proteins have been characterized (see review by Cassab and Varner, 1988; Varner and Lin, 1989). Extensins, the hydroxyproline-rich glycoproteins (HRGPs) of the dicotyledonous cell wall, are the best characterized cell wall proteins in higher plants (Chen and Varner, 1985b). Extensins have been found in many plant species and are proposed to be one of the major protein components of the primary wall (Lampert and Catt, 1981). The various levels of HRGP in different organs and tissues (Cassab et al., 1985; Cassab, 1986), however, suggest that the distribution of extensin in the plant cell is rather tissue specific or that HRGPs may not be the only major protein in plant cell walls. There are several known instances of total cell wall proteins in which glycine is a major amino acid component, suggesting the presence of glycine-rich protein (Varner and Cassab, 1986). Recently, glycine-rich protein genes were isolated from petunia and bean (Candit and Meagher, 1986; Keller et al., 1988). Additionally, cDNAs and genomic clones encoding proline/hydroxyproline-rich protein (PRP) distinct from the HRGP have been isolated from both soybean and carrot (Chen and Varner, 1985a; Hong et al., 1987). The isolation and characterization of genes representing these different classes of cell wall proteins have revealed that there are major differences in the basic repeat motifs of these proteins: Ser-(Hyp), for extensin, (Gly-X), for glycine-rich protein, and Pro-Pro-Val-X-Y for PRP.

We have recently reported the isolation and characterization of a proline-rich protein gene, ShbPRP1, of soybean, which probably plays an important role in plant development (Hong et al., 1987). The ShbPRP1 gene sequence predicted a novel proline-rich protein containing a highly repeated amino acid structure consisting essentially of Pro-Pro-Val-Tyr-Lys. The Northern blot analysis of RNA isolated from soybean hypocotyl tissue using a cDNA for this gene, pTU04, revealed that there are three mRNA bands homologous to the cDNA probe: a 1050 nt mRNA from the apical and elongating region, and two mRNAs of 1220 nt and 650 nt from the mature hypocotyl (Hong et al., 1987). Further studies on the expression of this PRP gene family using gene-specific probes revealed a highly regulated pattern of gene expression during plant development: the major form of PRP mRNA changed at different stages of development and/or in different tissues or organs (Hong et al., 1989).

To study the molecular basis of this developmental regulation of the ShbPRP gene family, we isolated and characterized two other PRP genes, ShbPRP2 and ShbPRP3, which encode the 1050 and 650 nt mRNAs, respectively. The results demonstrate that each member of this PRP gene family encodes a related proline-rich protein with differences occurring in...
the basicity and tyrosine content, as well as the size of the mature protein.

MATERIALS AND METHODS

cDNA Cloning of SbPRP2 mRNA—Ten μg of poly(A) RNA isolated from the apical region of the soybean hypocotyl (Hong et al., 1987) were used to make a cDNA library. cDNA libraries were then constructed by the method of Maniatis et al. (1982). The cDNA was treated with T4 DNA polymerase and the Klenow fragment of Escherichia coli DNA polymerase I (Maniatis et al., 1982) and methylated with EcoRI methylase prior to the addition of EcoRI linkers. Following digestion with excess EcoRI restriction enzyme and removal of excess linkers, the cDNA was size fractionated through a Sepharose CL-4B column to obtain cDNA molecules larger than 400 base pairs (Maniatis et al., 1982). The cDNA was then ligated into the unique EcoRI site in Xgt10 (Manners et al., 1989). Mungbean nuclei-protected DNA fragments were then obtained by the method of Maniatis et al. (1982). The DNA fragments were digested to determine the 5' and 3'-end mRNA termini as previously described (Hong et al., 1987). The 5'-end of 1050 nt mRNA was predicted from sequence comparisons to two other pTU04-related genes; the 3'-end of the RNA was deduced from the DNA sequence of several cDNAs which encode 1050 nt mRNA.

Construction of Gene-specific DNA Probes—The region of DNA sequence used as a gene-specific probe was determined. The fragment of the Charon 40 soybean genomic library were screened using a SbPRP2 gene-specific probe. Three X clones homologous to the SbPRP3 gene were isolated. The XJH403, which contained an internal HindIII site as mapped by restriction enzyme digestion. The region of DNA containing the gene-specific probe was identified from the DNA sequence to be centered in the 3'-untranslated region. Regions used for construction of probes were then subcloned into Bluescript plasmid.

RESULTS

Phage libraries constructed from soybean DNA were screened with pTU04 cDNA or pSAP4-5 cDNA as a probe (see “Materials and Methods”). Two genomic clones, AX41 and JH403, which carry the 650 nt RNA gene (SbPRP3) and the 1050 nt RNA gene (SbPRP2), respectively, were isolated and further characterized. Both genomic clones contained an entire copy of the respective gene flanked by several kilobases of DNA.

Nucleotide Sequence Analysis of the SbPRP2 and SbPRP3 Genes—The restriction maps and the nucleotide sequences of the SbPRP2 and SbPRP3 genes are shown in Figs. 1, 2 and 3, respectively. Restriction enzyme analyses and DNA sequence analyses of the different clones indicated that these two genes are not clustered. SbPRP2 contains an open reading frame of 690 bp (230 amino acids) whose deduced amino acid sequence encodes a 25,978 Da protein. This coding sequence was followed by a 3'-untranslated region that contained a consensuse sequence polyadenylation signal at 201 bp 3' to the translation termination codon. The SbPRP3 gene contains an open reading frame of 270 bp (90 amino acids) encoding a protein of 10,293 Da. A putative polyadenylation signal, AA-TAAA, is observed 236 bp 3' to the termination codon.

A summary of the transcription unit and encoded proteins, deduced from the nucleotide sequence of three SbPRP genes, is shown in Table I. From the deduced amino acid sequences, putative signal peptide sequences are predicted which are 22 and 23 amino acids in length for the SbPRP2 and the SbPRP3 genes, respectively. The patterns of N-terminal amino acid sequences are in good agreement with sequences which are typical of membrane spanning signal sequences of secretory proteins (Perlman and Halvorsen, 1983) (Fig. 7). Following the signal sequence, the remainder of the coding region consists of multiple 15-bp repeats.

Previously the SbPRP1 gene was reported to contain 43 repeats of a sequence consisting primarily of Pro-Pro-Val-Tyr-Lys (CCX-CXX-GTX-TAX-AAX). The SbPRP2 gene contains similar multiple 15-bp repeats. However, the deduced coding region reveals a protein composed primarily of 19 perfect alternating pentapeptide repeats: Pro-Pro-Val-Lys (CCX-CXX-GTX-TAX-AAX). These are five amino acid deviations from this repeat pattern, two Ile (ATT) substitutions for Val (GTT) and one substitution each of Thr (ACT) for Tyr (GTT), Asn (AAC) for Asp (GAD), and Gly (GGA) for Glu (GAG). The major difference between the SbPRP1 and SbPRP2 proteins is the Glu substitution for Tyr in the...
Proline-rich Cell Wall Protein Gene Family

FIG. 1. Structure and sequencing strategies for the SbPRP2 and SbPRP3 genes. The location of the gene and orientation of the transcription unit is shown as a thick arrow. The hatched boxes denote the regions which contain the DNA sequences used for the preparation of the gene-specific probes (GSP*).

A, restriction enzyme maps of the genomic clone, JH403, which carries the SbPRP2 gene (A), the subcloned fragments (B), and the cDNA pSAP4-5 (C) are illustrated. B, restriction maps of genomic clone, UL4x1, which carries the SbPRP3 gene (A), and a subcloned fragment pSAx41 (B) are illustrated. Structure and sequencing strategies of the SbPRP genes are illustrated in B. ATG and TGA denote the coding region. The enzyme sites are abbreviated as follows: AccI, A; Apd, Ap; BamHI, B; BglII, Bgl; EcoRI, E; HindIII, H; NdeI, Nd; PstI, P; PvuII, Pvu; ScaI, SC; SpeI, Sp.

The amino acid composition of the predicted mature SbPRP2 protein is Pro 39.1%, Lys 18.3%, Val 16.3%, Tyr 11%, and Glu 9.1% (Table I). The SbPRP2 protein is therefore predicted to be less basic than the SbPRP1 protein. The SbPRP3 gene also encodes a proline-rich protein having a pentapeptide repetitive structure which is shorter (90 amino acid) and more variable than that of the other two PRPs. Six repeats of Pro-Pro-Val-Tyr-Lys and two repeats of Pro-Pro-Tyr-Lys-Lys occur in SbPRP3.

The analysis of codon usage reveals a biased pattern with the preferred codon being CCA for Pro, GTT for Val, TAC for Tyr, and AAG for Lys.

Both SbPRP genes have putative CAT and TATA-motifs about -80 bp and -25 bp upstream of the cap sites, respectively. For SbPRP3 the 5'- and 3'-ends of the transcription unit were mapped using mungbean nuclease (data not shown) (Fig. 2). Since the SbPRP genes retain a high level of nucleotide sequence conservation and contain GTGTGTT prior to the transcription start site, the cap site of the SbPRP2 gene was deduced by analogy (Fig. 2). In SbPRP3, the upstream region of DNA contains two repeats of a 13-bp sequence (CATGCTTGATTt/aC) and two inverted repeats with a sequence of ATTGxxxCACTAxACATGCxA, but the significance of these conserved sequences has not been determined.

Sequence Homology of the SbPRP Gene Family—A diagram of the organization of the SbPRP gene family of soybean is illustrated in Fig. 5. The SbPRP1 gene was included for comparison. These genes show striking similarity with several regions of sequence identity observed at similar locations in the 5'-upstream regions around the CAT and TATA motifs: the 5'-untranslated sequences, the coding regions, including a signal peptide sequence and the mature protein, and the 3'-untranslated region around the polyadenylation signal.

In the 5'-upstream region of the three SbPRP genes, nucleotide sequence conservation extends into the putative CAT boxes, GTCAx₅₋₇T, which is positioned approximately the same distance from the cap site (~85 bp) upstream of the cap sites. The nucleotide conservation ranges from 69% (between SbPRP2 and SbPRP3 or SbPRP1 and SbPRP3) to 84% (SbPRP1 and SbPRP2) in this 5'-conserved region. In addition, immediately 5' to the transcription initiation site, a GTGTGTT sequence is present in all three SbPRP genes. When the regions of DNA encoding the putative signal sequences are compared, it is apparent that these PRPs are closely related. The sequence homologies are:

- SbPRP1 and SbPRP3
- SbPRP1 and SbPRP2
- SbPRP2 and SbPRP3

These PRPs are closely related.
### TABLE I

Comparison of three proline-rich protein genes of soybean

<table>
<thead>
<tr>
<th>Gene</th>
<th>Size of transcripts</th>
<th>Transcription unit</th>
<th>Encoded protein</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(nt)</td>
<td>(bp)</td>
<td>(aa)</td>
</tr>
<tr>
<td>SbPRP1</td>
<td>1,220</td>
<td>1,076</td>
<td>256</td>
</tr>
<tr>
<td>SbPRP2</td>
<td>1,050</td>
<td>947</td>
<td>230</td>
</tr>
<tr>
<td>SbPRP3</td>
<td>650</td>
<td>768</td>
<td>208</td>
</tr>
</tbody>
</table>

#### Notes

- **Exon** (bp) 1,076 | 947 | 768 | 650 |
- **5'-UNT** 20 (297) | 28 | 255 | 270 |
- **3'-UNT** 285 | 226 | 255 |
- **ORF** 768 | 690 | 270 |

### Fig. 3. The nucleotide sequence and deduced amino acid sequence of the SbPRP3 gene.

The transcribed region of DNA is denoted as mRNA (+ 1 to bold arrowhead). The putative CAT box, TATA motif, and poly(A) addition signal are boxed. The amino acid repeat, Pro-Pro-Val-Tyr-Lys is single underlined. The 13-bp direct repeat (A) and the imperfect inverted repeat (B) are double underlined. Nucleotides are numbered with the transcription initiation site as +1.

### Fig. 4. Southern blot analysis of soybean genomic DNA hybridized with the 3'-noncoding region gene-specific probes of SbPRP2 and SbPRP3.

Ten μg of soybean DNA were digested with the indicated restriction enzymes, electrophoresed on a 0.6% agarose gel, and transferred to nitrocellulose. Indicated molecular sizes in kilobases were determined by HindIII-digested DNA molecular mass markers.

### Fig. 5. Schematic diagram of the structural organization of three SbPRP cell wall genes of soybean.

Regions of high sequence homology among the SbPRP genes are boxed; regions of no similarity are indicated with bold lines. The amino acid repeat structure for each gene is indicated. The region of DNA from which gene-specific probes were obtained is indicated by GSP.*

### Fig. 6. Comparison of the N-terminal amino acid sequence of three SbPRPs.

A signal peptide sequence and the putative cleavage site is indicated. Boxed regions indicate the amino acid sequences which form the hydrophobic core of signal peptide sequences. The putative cleavage site was predicted according to Von Heijne (1984).

### Notes

- **SbPRP2** 7.7 Kb
- **SbPRP3** 5.4 Kb
- **SbPRP1** 3.6 Kb
- **SbPRP3** 3.6 Kb

---

*Hong et al., 1987.*
protein genes, SbPRP2 and SbPRP3, which encode 1050 and 650 nt mRNAs, respectively, reveal a striking similarity among members of this gene family, detectable both at the nucleotide sequence level and in the highly repetitive structure observed for the deduced amino acid sequences of polypeptides encoded by all three SbPRP genes (Fig. 5). The presence of highly repetitive structural motifs is explained by the cross-hybridization of the multiple 15-bp repeats (CCX-CCX-GTX-TAX-AAX) in all mature SbPRP3 protein, which is approximately 8 kDa versus 1050 and 650 nucleotide mRNAs on North-Westley cDNA to 1050 and 650 nucleotide mRNAs on North-Westley cDNA to 1050 and 650 nucleotide mRNAs.

In a recent study, the steady-state levels of SbPRP mRNA accumulation showed that each transcript is differentially expressed, revealing dramatic patterns of developmental and organ specificity (Hong et al., 1989). The major SbPRP transcript accumulating in the mature hypocotyl, root, and young seed coat was SbPRP1. SbPRP2 in the apical hypocotyl and young tissue cultured cells, and SbPRP3 in most of the aerial parts of the soybean plant.

In contrast to the marked differences in mRNA accumulation, sequence analysis of the SbPRP genes revealed strikingly similar gene structure and a high degree of sequence conservation. The nucleotide sequence conservation extends into the 90 to –100 bp region upstream of the cap site and 3′ into the noncoding region around the polyadenylation signal. The DNA sequences around the CAT and TATA motifs, cap sites, and polyadenylation signal are positioned at approximately the same distances relative to the transcription unit. This high conservation suggests functional significance of these sequences for the regulation of expression of the SbPRP gene family. However, sequences involved in the dramatic developmental- and organ-specific expression of this gene family are likely located upstream of this highly conserved region.

To address the functional importance of the proline-rich protein gene family, it is desirable to find a relationship to other known cell wall protein genes. The most highly studied cell wall proteins are the HRGP extensins of dicots (reviewed by Cassab and Varner, 1988). Cell wall HRGP extensins contain a characteristic pentapeptide repeat sequence, the Ser-(Pro)₄N, in the primary translation product of the apoprotein, which is subsequently hydroxylated to give Ser-(Hypro), found in the mature protein (Chrispeels, 1970). A comparison of amino acid sequences deduced from all known HRGP sequences revealed substantial variability in the length or in amino acids flanking the highly conserved Ser-Pro repeats, e.g. Val His or Val Ala (Showalter et al., 1985), Thr-Pro-Val-Tyr-Lys (Smith et al., 1986), and Val-Tyr-Tyr-Tyr-Lys or Tyr-Tyr-Tyr-His (Corbin et al., 1987) (see review by Showalter and Varner, 1989). These flanking sequences are presumed to be important in functionally and structurally distinguish ing each HRGP isomer. Cell wall HRGPs contain carbohydrates, approximately two thirds of the glycoprotein mass, which are composed largely of oligosaccharinolyl residues attached through O-glycosidic linkages to most of the hydroxyproline residues and, to a much lesser extent, by galactose, which occurs in U-glycosidic linkage to some of the serine residues (Lampert, 1973). Glycosylation of extensin has been suggested to be important in maintaining its proper structure and function (Stafstrom and Staehelin, 1986; Sadava and Chrispeels, 1973; Cooper et al., 1983). The polysaccharides of mature HRGP appear to stabilize HRGP in forming extended helical rods of polyproline II conformation (left-handed helix, 3 residues/turn, a pitch of 3.12 Å) (Van Hulse and Varner, 1984). The HRGP becomes insoluble in the cell wall, perhaps through the formation of isodityrosine bonds (Fry, 1986), which are formed by two adjacent tyrosine residues. Although Ser-Pro does not occur in the SbPRP gene family, the relationship of SbPRPs to other cell wall HRGPs is noted at the nucleotide sequence level (Hong et al., 1987). The Pro-Pro-Val-Tyr-Lys repeat of SbPRPs has been observed as part of an extensin repeat in tomato (Smith et al., 1986). Recently, additional cDNAs encoding cell wall proteins have been reported which are unrelated to the HRGP and lack the Ser-Pro motif. These include cDNAs for nodulin-75 (Frassen et al., 1987) of soybean nodule and a 33 kDa PRP of carrot (Chen and Varner, 1985a).

A cell wall protein which has the same amino acid composition as the SbPRP1 protein was identified from cultured soybean cells (Averyhart-Fullard et al., 1988). The protein lacks histidine and serine and contains 20% hydroxyproline and 20% proline. The molecular mass of the protein appears to be 33 kDa on sodium dodecyl sulfate-polyacrylamide gel.
which is higher than the value of the 27 kDa mature SbPRP1 protein predicted from the nucleotide sequence. However, the presence of repetitive proline-rich fragments and hydroxyl-protein predicted from the nucleotide sequence. If SbPRPs occur as non-glycosylated or slightly glycosylated cell wall protein in soybean, this family of protein would provide different function than the HRGP extensins. The high tyrosine content in all three SbPRPs implicates the high tyrosine content in all three SbPRPs implicates the 1CH35 and ACH40 soybean genomic libraries. 

It is of interest to note that a cell wall protein of maize, a Graminaceous monocot, contains a lower proportion of hydroxyproline and a lower carbohydrate content compared with dicot HRGP (Lampart and Miller, 1971; Kieliszewski and Lampart, 1987; Hood et al., 1988). The nucleotide sequence of a cDNA clone encoding a maize cell wall protein revealed a lack of Ser-(Pro), as a major repeat structure since this sequence occurs only once in the coding region (Stiefel et al., 1988). In this respect SbPRPs are rather similar to the cell wall protein of this monocot than to the well described HRGPs of dicots.

The study of SbPRP gene expression showed that PRP genes are highly expressed in the root, seed coat, and tissue-specific regulation of the PRP cell wall protein gene family. The striking similarity of these genes is not only seen in the amino acid sequences of the predicted proteins but also by the regions of high DNA sequence homology in both some 5' and 3'-noncoding regions. The functional difference and the selective expression of the SbPRPs, however, is not apparent from sequence comparisons. Experiments to localize tissue- and developmental-specific expression will help address this question.

Acknowledgments—We thank Drs. J. Slightom and Yu Ma for providing the XCH35 and XCH40 soybean genomic libraries.

REFERENCES


J C Hong, R T Nagao and J L Key


Access the most updated version of this article at http://www.jbc.org/content/265/5/2470

Alerts:
• When this article is cited
• When a correction for this article is posted

Click here to choose from all of JBC's e-mail alerts

This article cites 0 references, 0 of which can be accessed free at http://www.jbc.org/content/265/5/2470.full.html#ref-list-1