Primary Structure of 6.5k-Arginine/Glutamate-rich Polypeptide from the Seeds of Sponge Gourd (Luffa cylindrica)

Makoto Kimura,* Sung-Soo Park,* Ritsu Sakai,* Nobuyuki Yamasaki,* and Gunki Funatsu**

*Laboratory of Biochemistry, Faculty of Agriculture, and **Laboratory of Protein Chemistry and Engineering, Faculty of Agriculture, Kyushu University, Fukuoka 812–81, Japan
Received November 20, 1996

The amino acid sequence of 6.5k-arginine/glutamate rich polypeptide (6.5k-AGRP) from the seeds of sponge gourd (Luffa cylindrica) has been determined. The 6.5k-AGRP consists of a 47-residue polypeptide chain containing two disulfide bonds, and a molecular mass calculated to be 5695 Da, which fully coincides with a value of [M+H]+ = m/z 5693.39 obtained by matrix-assisted laser desorption ionization time of flight mass spectrometry (MALDI-TOF MS). The mass spectrometric evidence indicated that 6.5k-AGRP is also present partially truncated at the C-terminus. In our preparations, approximately half of the polypeptide molecules have the C-terminal sequence Arg-Arg-Glu-Val-Asp; the other half lack Val-Asp and end with the glutamic acid, making a total of 45 residues in the polypeptide chain. The two disulfide bonds connect Cys12 to Cys23 and Cys12 to Cys39. Comparison of the amino acid sequence of 6.5k-AGRP with those of the other known proteins included in the PIR protein sequence database showed that it is related to the amino acid sequence of the N-terminal region encoded by the first exon of the cocoa (Theobroma cacao) and cotton seeds vicilin genes, sharing a characteristic two Cys-Xaa-Xaa-Cys motif.

Key words: amino acid sequence; sponge gourd; Luffa cylindrica; storage protein; arginine/glutamate-rich polypeptide

Plant seeds contain large amounts of storage proteins that provide N and C reserves during germination and seedling growth. Because synthesis of storage proteins is highly regulated in both time and place, they are expressed as large precursors, undergo co- and post-translational modification, and are transported into discrete deposits called protein bodies, they have provided an interesting experimental system with which to study gene regulation, protein processing, and compartmentalization.3,4 Information that suggests the role as a regulatory protein for their own genes has been accumulated on storage proteins. That is, napin, which is a 2S albumin from Brassica napus contains a short track rich in basic amino acid, typical of nuclear localization signals,5 and also a glutamine-rich domain that are characteristic of one group of transcription factors.3,4 Furthermore, the three dimensional analysis of phasolin, another seeds storage protein from Phaseolus vulgaris, revealed a domain with structural similarity to the helix-turn-helix motif found in certain DNA binding proteins.5

Previously, Ishihara et al. isolated four arginine/glutamate-rich polypeptides (AGRPs) with different molecular sizes, ranging from 5kDa to 14 kDa from the sponge gourd (Luffa cylindrica) seeds and analyzed them with respect to amino acid composition, N-terminal sequence, and localization in protein bodies.6 To examine the function(s) of AGRPs in the sponge gourd seeds, we have sought to establish their detailed primary structures. In this study, we determined the complete amino acid sequence of 6.5k-AGRP and discuss its sequence similarity to other known protein sequences.

Materials and Methods

Materials. Seeds of the sponge gourd were obtained from Takii Seed Co., Ltd. 6.5k-AGRP was purified by the procedures described in our previous paper6 except that ion-exchange column chromatography on a CM-cellulose column was omitted. In brief, the pH 4.0 extract from the sponge gourd seeds was put onto a Sephadex G-75 column (5 x 140 cm) and eluted with 50 mM sodium acetate buffer, pH 5.0. The fractions containing 6.5k- and 12.5k-AGRPs were pooled, dialyzed against 10 mM phosphate buffer, pH 7.0, containing 3.0 mM ammonium sulfate, and put onto a butyl-Toyopearl 650M column (1.5 x 14 cm) previously equilibrated with the same buffer. The polypeptides were eluted with a linear gradient of ammonium sulfate from 3 to 0 M in 200 ml of 10 mM phosphate buffer, pH 7.0. In this chromatography, 6.5k-AGRP was eluted at 1.2 M ammonium sulfate concentration, and it gave a single peak in reverse-phase HPLC (RP-HPLC) on Wakosil C18 column (4.6 x 150 mm) and a single band in tricine SDS-PAGE (data not shown). Lysylendopeptidase and TPCK-treated trypsin were purchased from Wako Pure Chemicals and Sigma Chemical Co., respectively. All reagents were of analytical grades.

Enzymatic digestsions and separation of peptides. One hundred μg of the S-pyridylethylated 6.5k-AGRP, prepared by the procedure of Cavins and Friedman,7 was dissolved in 100 μl of 0.2 M N-methylmorpholine acetate buffer, pH 8.1, and digested with 2 μg of lysylendopeptidase at 37°C for 2 h. For placement of the disulfide bonds, the intact 6.5k-AGRP was digested with trypsin using an enzyme/substrate ratio of 1 to 30 at 37°C for 6 h.

The peptides derived from enzymatic digestions were separated by RP-HPLC on either a Wakosil C18 column (4.6 x 150 mm) or YMC gel C4 column (4.6 x 250 mm) using an acetonitrile gradient in 0.1% aqueous trifluoroacetic acid (TFA) at a flow rate of 1.0 ml/min. The elute was monitored by absorption at 220 nm.

Amino acid analysis. Peptides (1 nmol) were hydrolyzed in the vapor of 5.7 N HCl containing 0.02% 2-mercaptoethanol at 110°C for 24 h in evacuated sealed tubes. The amino acids were analyzed on a Shimadzu LC6A system amino acid analyzer after derivatization of amino acids with phenylisothiocyanate, as described by Henrikson and Meredith.9 The amino acids were sequenced by a gas-protein sequencer, PSQ-1 (Shimadzu). Phenylthiohydantoin (PTH) amino acids were separated by RP-HPLC on a Wakosil PTH (4.6 x 250 mm) with an isocratic system.

Mass spectrometry. The molecular weights of the intact 6.5k-AGRP and of peptides were measured on a Voyager matrix-assisted laser desorption ionization time of flight mass spectrometry (MALDI-TOF MS) (PerSeptive Biosystems) with a 337-nm, N2 laser and an ion reflector. The spectra of
polypeptides were acquired in the positive-ion linear mode and with delayed ion extraction at an accelerating potential of 30 kV. α-Cyano-4-hydroxycinnamic acid in acetonitrile-water-0.1% TFA (50:40:10) was used as the matrix and insulin was used for external calibration.

**Homology search.** A homology search through the PIR (Protein Identification Resource), release 49.0 June 1996 protein database of protein sequence, was done by the program BLAST.⁶

**Results and Discussion**

**Isolation**

6.5k-AGRP was isolated by gel filtration on a Sephadex G-75 column followed by butyl-Toyopearl 650M column chromatography, as described under Materials and Methods. The polypeptide thus obtained provided a single N-terminal amino acid sequence: Pro-Arg-Gly, which was identical to the result reported previously. In contrast to the sequence result, the mass spectrometry analysis of the polypeptide yielded two separated peaks, showing values of [(M+H)⁺ =m/z 5479.48 and 5693.39 (Fig. 1). The result indicated that the 6.5k-AGRP prepared in this study was a mixture of two isoforms with the identical N-terminal amino acid sequences. However, no attempt to separate two isoforms was made in this study and the polypeptides were directly subjected to sequence analysis without prior separation into two components.

**Sequence analysis**

In a previous paper, the N-terminal 20 amino acid sequence of 6.5k-AGRP was reported. In this study, N-terminal amino acids were again thoroughly sequenced as far as possible. This analysis unambiguously identified 39 amino acid residues, as shown in Fig. 2, and no microsequence heterogeneity was found at individual positions.

Since the amino acid analysis showed the presence of one lysine residue in 6.5k-AGRP, which was identified at the position 35 by direct sequencing, lysylendopeptidase was chosen as an enzyme for fragmentation of the polypeptide. S-Pyridylethylated 6.5k-AGRP was digested with lysylendopeptidase, as described under Materials and Methods, and the digests were separated by RP-HPLC (Fig. 3A). This digestion yielded three peptides K1, K2, and K3, and they were characterized with respect to amino acid composition, sequence, and molecular mass. Sequencing and amino acid compositional analyses identified peptide K3 to be the N-terminal peptide from the position 1 to 35, K1 from 36 to 45, and K2 from 36 to 47. This result indicated that a pair of peaks found in mass spectrometry analysis was due to the presence of the C-terminally truncated form: the longer molecule ends with Glu, while the shorter with Asp.

The C-terminal truncation was further corroborated by mass spectrometric analysis of peptides K1 and K2, showing the [M+H]⁺ ions of m/z 1355.08 and 1571.65, respectively, as shown in Fig. 3B. The mass difference (216.57 Da) of peptides K1 and K2 matched well with the difference of the [M+H]⁺ ions (m/z 213.91) observed by mass spectrometric analysis of the two components of the 6.5k-AGRP preparation described above. These results led to
the conclusion that the difference between the two components in the preparation of 6.5k-AGRP must be the only C-terminal truncation, and no another isoforms of 6.5k-AGRP was present in the preparation. In this way, the complete amino acid sequence of 6.5k-AGRP was analyzed as shown in Fig. 2.

Placement of disulfide bonds

The amino acid analysis showed the presence of four cysteines in 6.5k-AGRP. This study identified them at positions 12, 16, 29, and 33. To define the disulfide bonds in the polypeptide, the intact 6.5k-AGRP was digested with trypsin and the resulting peptides, which were separated by RP-HPLC (Fig. 4), were analyzed with respect to amino acid composition. The result localized tryptic peptides T1, T2, and T3, as shown in Fig. 2. Also it showed that only peptide T4, which yielded a single peak in mass spectrometry (data not shown), contained cystine residues and its amino acid composition was in good agreement with the sum of the amino acid residues of peptides, positions 7–13, 16–25, and 29–35. This result suggested that three tryptic peptides, Thr12–Arg15, Cys16–Arg33, and Cys39–Lys35, might be connected with each other by disulfide bonds, as shown in Fig. 4. It was therefore likely that the two cystine residues (Cys39 and Cys35) in tryptic peptide at position 29–35 may not be connected, but rather either Cys12 may be connected to Cys29 and Cys16 to Cys33, or Cys12 may be linked by Cys33 and Cys16 by Cys29, as shown in Fig. 4. The PTH-cystine is known to be separated by the other PTH amino acids, being eluted at around 8 min from a Wakosil PTH column (4.6 × 250 mm) with an isocratic system. Hence, the peptide T4 was directly sequenced by a gas-phase sequencer. Figure 5 shows RP-HPLC patterns for identification of PTH amino acids obtained by each Edman degradation step. The analysis clearly identified PTH-cystines at the first and sixth steps. Hence, we conclude that the two disulfide bonds in 6.5k-AGRP connect Cys12 to Cys33 and Cys16 to Cys29.

Characterization

The amino acid composition derived from the sequence is Asp1, Thr1, Ser1, Glu8, Gln4, Pro2, Gly3, Ala2, Val3, Leu2, Tyr1, Lys1, His1, Arg2, and 1/2Cys4. This agrees well with the result obtained from the amino acid analysis of the polypeptide. The amino acid composition of 6.5k-AGRP gives 5695 as its M, value, which is in good agreement with the value of [M+H]+ = m/z 5693.39 obtained by MALDI-TOF MS. As expected, arginine and glutamic acid are the most abundant amino acids in 6.5k-AGRP, contributing 25% and 17%, respectively, of the total residues in the polypeptide. They are distributed throughout the sequence but there is a tendency for them to occur in small clusters in the C-terminal half of the molecule.

It has been reported that there are usually a number of closely related storage proteins in seeds, which are produced by expression of a gene family. This study, however, showed that the storage protein 6.5k-AGRP from the sponge gourd might be encoded by a single type of the gene. It is also known that most of the storage proteins undergo
reported that in addition to other vicilin-like sequences, sucrose binding protein: monomeric 62 kDa protein isolated from Glycine max,\textsuperscript{13} and trypsin inhibitors (6 kDa monomeric proteins) from buckwheat seeds\textsuperscript{14} are related to the vicilin sequences. Hence, 6.5k-AGRP is aligned with these similar proteins for maximum similarity, as shown in Fig. 6A. 6.5k-AGRP can be easily aligned with the N-terminal hydrophilic segment known as a “insert” in two vicilins and sucrose binding protein, and also with the entire length of buckwheat trypsin inhibitor. The characteristic two Cys-Xaa-Xaa-Xaa-Cys motif in a “insert” are absolutely conserved in all molecules (positions 12–16 and 29–33 in the 6.5k-AGRP sequence) and at a considerable number of the other positions there are chemically similar residues. It is therefore speculated that the ancestral proteogene encoding 6.5k-AGRP and/or buckwheat trypsin inhibitor had been derived from a “insert” in the first exon of the vicilin genes. Alternatively, perhaps more likely, a “insert” in the first exon of the vicilin genes have occurred by insertion of the ancestral 6.5k-AGRP-like proteogene into the first exon during evolution of the vicilin gene; this event might have given rise to some vicilins that reserve more extensive N, C, and S resources than other vicilins without a “insert” in the N-terminal region.

As described in introduction, it has been reported recently that some storage proteins have a characteristic amino acid sequence or structure which might be involved in transcriptional regulation. Interestingly, the similarity search also showed that the sequence of 6.5k-AGRP is significantly similar to U1 snRNP 70k protein\textsuperscript{15} and transcriptional factor DP-2,\textsuperscript{16} as given in Fig. 6B. The significance of this relatedness must await more detailed studies on 6.5k-AGRP.

Sequence comparison

It was suggested from a comparison from the N-terminal 20 amino acid sequence of 6.5k-AGRP\textsuperscript{16} that it might be related to that of cocoa (Theobroma cacao) seed vicilin.\textsuperscript{11} After analysis of the complete amino acid sequence of 6.5k-AGRP, we compared its structure with the other known protein sequences using the computer program BLAST. This analysis again showed that 6.5k-AGRP is similar to the hydrophilic N-terminal sequence of the cocoa vicilin, and also to that of cotton vicilin.\textsuperscript{12} Although 6.5k-AGRP significantly differs from them in a size. Vicilins are typically trimeric proteins of \( M \), 150,000 to 190,000; their subunits are polypeptides of \( M \), 60,000 and they are a widely distributed group of storage proteins not only in angiosperms but also in gymnosperms. Recently, it was

---

References

Fig. 5. Identification of PTH-amino Acids by RP-HPLC on a Wakosil PTH Column.

PTH-amino acids were separated on Wakosil PTH column (4.6 × 250 mm) and eluted with PTH-amino acids mobile phase (Wako Pure Chemicals). DPTU indicates diphenylthiourea.

A

6.5k-AGRP
PGGSPRTETYRSACGRQRQWAEAGVER25
Cov
YERDRPQCQYRCQBCVSSZELHRRR58
BWI-2a
SDYRQLOEQCCNNCLTREWSTDN24
Sbp
TVEYREDPVLTQHQLCQQQYTED57

6.5k-AGRP
QRCCQVCEXRLBEQRGRRVBD47
Cov
GEOCQQRGCRHRYQQRQES80
BWI-2a
QRCQQRQGQQRQ123
Sbp
KRCQEGQDKSRYHHQKEQ78

Fig. 6. Comparison of the Amino Acid Sequence of 6.5k-AGRP with Those of the Related Proteins.

A: Alignment of 6.5k-AGRP with the other vicilin related proteins, including coca vicilin (Cov)1,2 cotton vicilin (Cov)-1,2,1 buckwheat trypsin inhibitor (BWI-2a),2,15 and sucrose binding protein (Sbp)1,2. Amino acids identical to those of 6.5k-AGRP are boxed.

B: Alignment of 6.5k-AGRP with U1 snRNP 70k protein (U1 snRNP)2,15 and transcriptional factor DP-2 (TF-DP-2).15,21 Identical residues are boxed.