Volume 13 Number 18 1985 Nucleic Acids Research

A cDNA clone encoding a glycinin A_{1a} subunit precursor of soybean

Takaharu Negoro, Takayuki Momma and Chikafusa Fukazawa*

Genetic Engineering Laboratory, Division of Applied Microbiology, National Food Research Institute, Ministry of Agriculture, Forestry and Fisheries, Tsukuba, Ibaraki 305, Japan

Received 2 July 1985; Revised and Accepted 6 September 1985

ABSTRACT

A cDNA clone covering the whole coding region for a glycinin subunit precursor containing the A_{1a} acidic subunit, one of the A_2 family, has been identified from a library of soybean cotyledonary cDNA clones using a mixed oligonucleotide probe. Analysis of the cDNA insert revealed that it contained 1746 nucleotides of mRNA sequence with a 5'-terminal nontranslated region of 54 nucleotides, a signal peptide region corresponding to 19 amino acids, an acidic subunit region (A_{1a}) corresponding to 291 amino acids followed by a basic subunit region corresponding to 185 amino acids, and a 3'-terminal nontranslated region of 207 nucleotides. By comparing the predicted protein sequence of this precursor with that of the legumin A precursor of pea, it was found that glycinin A_2 subunit family appeared to be more closely related to the legumin than to the A_3 subunit family, and that the evolutional rearrangement of glycinin genes has occurred.

INTRODUCTION

Glycinin, the principal seed reserve protein of soybean (Glycine max L., Merr.) is a hexameric molecule which has a sedimentation coefficient ($S_{20,w}$) of $\sim$ 12 and a molecular weight (Mr) of $\sim$ 360,000. It is composed of six subunit pairs, each of which comprises an acidic polypeptide component (Mr 35,000-42,000) which is disulfide-linked to a basic polypeptide component (Mr $\sim$ 22,000) (1-3). There are several specific acidic/basic subunit pairs, since six different acidic and five basic components were identified on the basis of differences in the partial primary structure of their NH_2-terminal regions (4). These subunit pairings of glycinin molecules are not random but heterogeneous, and all subunits except A_4 have been determined to derive from the covalently linked acidic/basic subunit pairs $A_{1a}B_2$, $A_{1b}B_{1b}$, A_2B_{1a}, A_3B_4 and A_5B_3 (5).

More recently, the amino acid sequences derived from the cDNA sequences for the A_3B_4 and the $A_5A_4B_3$ (6,7), and from the partial genomic DNA sequence for the A_2B_{1a} (8) have confirmed that these specific subunit complexes are synthesized from the same mRNA as a single precursor polypeptide which is cleaved to form a unique disulfide-linked subunit pair during post-translational processing as speculated previously (9). In order to understand the gene number, the evolutionary divergence and the level of transcription for every glycinin subunit precursor, it is necessary to determine the cDNA coding sequence of each gene.

In this report, we present the nucleotide and the deduced amino acid sequence of a glycinin subunit precursor including the A_{1a} acidic polypeptide component.

EXPERIMENTAL PROCEDURES

Preparation of RNA and construction of a cDNA library

Total RNA was prepared from the developing cotyledonary tissue (38 days after flowering) of soybean (*G. max* L., Merr. cv. Bonminori) as described previously (6) and poly(A)-containing RNA was obtained by oligo (dT)-cellulose affinity chromatography (10). Double-stranded cDNA was synthesized from the above total poly(A)-containing RNA, sized, tailed, inserted and then transfected into *Escherichia coli* RR1 or *E. coli* GM2150 as described previously (6,7). Tetracycline-resistant and ampicillin-sensitive transformants were screened by colony DNA-filter hybridization to ^{32}P-labeled cDNA prepared from the total poly(A)-containing RNA. Those clones hybridizing to the [^{32}P]cDNA were selected for further screening with the specific oligonucleotide probe as described below.

Preparation of oligonucleotide probe

A synthetic oligonucleotide probe used to select the cDNA clones containing the nucleotide sequence corresponding to the NH_2-terminal unique sequence of the glycinin A_2 subunit family was synthesized by the solid-phase phosphite method (11), using morphilinoamidites, and then purified by reverse-phase high-preformance liquid chromatography. The probe, constructed to correspond to all possible codons for the hexapeptide Gln-Gln-

Asn-Glu-Cys-Gln (residues 6-11), was a 32 fold degenerate 17 mer (the third base for Gln-11 being excluded) and the sequence was 3'-GT$^{T}_{C}$GT$^{T}_{C}$TT$^{A}_{G}$CT$^{T}_{C}$AC$^{A}_{G}$GT.

Screening of soybean cotyledonary cDNA library

The library was screened with the ^{32}P-labeled oligonucleotide according to the procedure described previously (6).

Plasmid DNA preparation and restriction analysis

Plasmid DNA was isolated by the alkaline lysis method (12) and further purified on CsCl equilibrium density gradients. Restriction enzymes were purchased from Nippon Gene Inc., Toyobo Biochemicals and New England Biolabs and used as recommended by the vendor.

Nucleotide sequencing

The cDNA inserts of fourteen positive clones were sequenced partially or completely by the method of Maxam and Gilbert (13) and examined by computer analysis (Software Development Corporation, Japan).

RESULTS

Nucleotide sequence of cloned cDNA encoding A_{1a} acidic subunit

About 2,000 clones were screened with the oligonucleotide probe; eighty-three gave positive signals with this probe. These positive clones, which might cover the NH_2-terminal region of the glycinin A_2 subunit family, i.e. the A_1 and the A_2 acidic subunits, were classified into two groups on the basis of the restriction patterns for the inserts. Some of each group were selected in order to determine the partial nucleotide sequences. As a result of these nucleotide sequence analyses and a comparison of their deduced amino acid sequences with the NH_2-terminal sequences of glycinin subunits (references 4 and 5), a recombinant plasmid which contained a cDNA insert of 1746 nucleotides was selected as one of the longest plasmids containing NH_2-terminal region of the A_{1a} acidic subunit.

The complete nucleotide sequence of this cloned glycinin subunit cDNA was determined. The restriction endonuclease cleavage fragments and the sequencing strategy used in the nucleotide sequence determination are indicated in Fig. 1.

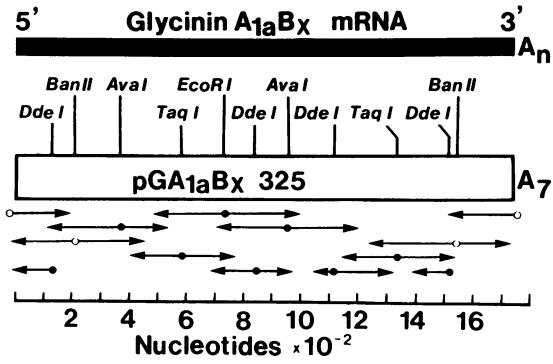

Figure 1: Restriction endonuclease map and nucleotide sequencing strategy for glycinin $A_{1a}B_x$ subunit ds-cDNA insert of $pGA_{1a}B_x$ 325. The top horizontal line represents the length of $A_{1a}B_x$ mRNA corresponding to the relative positions of the cloned ds-cDNA. The cloned ds-cDNA is shown with only those restriction endonuclease sites employed in the sequence determination. Each of the ends of the ds-cDNA is bound by segments containing about 16 G.C residues and also contains a coding-strand oligo (dA) segment of seven nucleotides at the 3' end. Arrows indicate the direction and length of sequencing of restriction endonuclease-digested DNA fragments. The solid circles represent the sites labeled at recessed 5' ends and the open circles represent the sites labeled at protruding 3' ends.

Figure 2 shows the nucleotide sequence of the insert, which corresponds to the entire coding region of the mRNA and adjacent portions of its untranslated sequences. According to a sequence definition proposed by Turmer et al. (4), the deduced subunit complex denotes $A_{1a}B_{1b}$. However, we have avoided naming the sequence because we have not been convinced that this clone really is talking about $A_{1a}B_{1b}$ rather than any other member of that closely related group. Therefore, this clone was designated as $A_{1a}B_x$. At the 5'-terminus a nontranslated region of 54 nucleotides, preceding the AUG translation start codon, was found in the cloned cDNA, whereas the position of the NH_2-terminal phenylalanine residue of mature A_{1a} subunit (with its codon beginning at position 112 in the nucleotide sequence shown in Fig. 2) was identified by a comparison to the partial NH_2-terminal sequence of the corresponding polypeptide (4). The 19 codons, preceding this NH_2-terminal phenylalanyl codon of the A_{1a} subunit, appear to encode a signal peptide, which is rich in hydrophobic amino acids and contains a basic amino acid such as

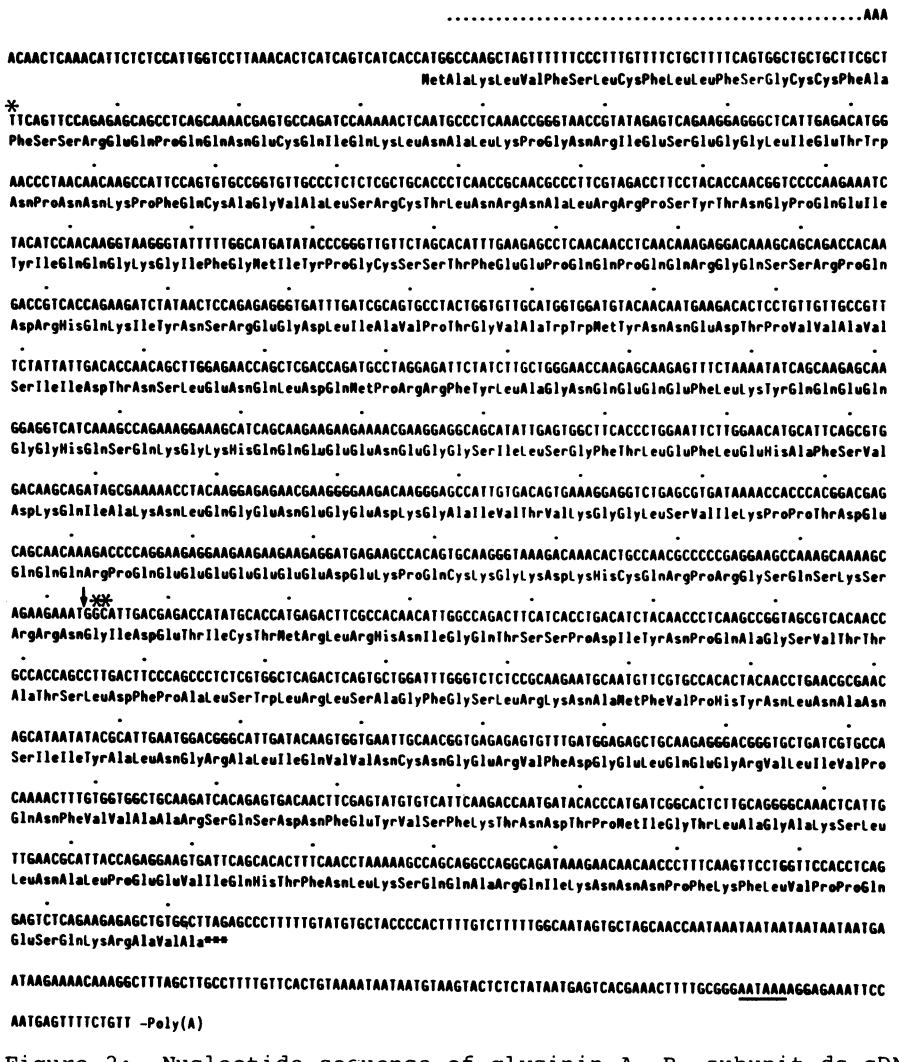

Figure 2: Nucleotide sequence of glycinin $A_{1a}B_x$ subunit ds-cDNA. The complete nucleotide sequence for the coding strand of the cloned ds-cDNA is shown with the predicted amino acid sequence for the primary translation product of glycinin $A_{1a}B_x$ subunit mRNA. A single asterisk and the two asterisks indicate the NH_2-terminal residues of the mature A_{1a} and the B_x subunit, respectively. The cleavage site of the subunit precursor is indicated by the arrow and the termination codon is at the three asterisks. A putative polyadenylation signal sequence is underlined.

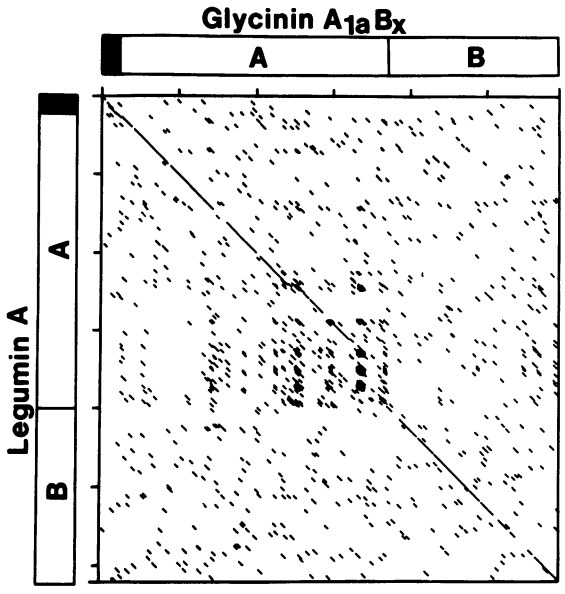

Figure 3: Dot matrix comparison of the coding regions in the mRNAs encoding glycinin $A_{1a}B_x$ subunit and legumin A subunit precursors. In the sequence comparisons plotted here, each dot corresponds to the first nucleotide position in a series of ten for which at least eight nucleotides match. The diagonal of the plot represents a base-to-base alignment of the two mRNA sequences, whereas the diagonal strings of dots which lie on either side of the center diagonal correspond with homologous sets of nucleotides. The coding regions of both mRNAs are designated at the top and left of the figure, respectively. The solid boxes represent the nucleotide sequences corresponding to the signal peptide segments. The regions designated as A indicate the nucleotide sequences corresponding to the acidic polypeptide components, while the regions designated as B represent the sequences corresponding to the basic polypeptide components of both precursors.

lysine prior to this hydrophobic segment. Analysis of the coding region of glycinin $A_{1a}B_x$ subunit mRNA also indicated that the A_{1a} acidic and the B_x basic subunits were synthesized as a single precursor polypeptide, which underwent post-translational processing to form an A_{1a}-B_x intermediary subunit complex, covalently linked by a disulfide bond. The mature A_{1a} subunit is encoded by 873 nucleotides (291 amino acids) and the mature B_x subunit is by 555 nucleotides (185 amino acids) in the mRNA. The relative molecular masses, calculated from the deduced amino acid

sequences of A_{1a} and B_x subunits, were 33032 and 20408, respectively. Following the UAG translation stop codon is a 3'-terminal nontranslated region of 207 nucleotides, adjacent to a poly(A) tail. The sequence AAUAAA, a putative polyadenylation signal sequence (14), is located in the glycinin $A_{1a}B_x$ mRNA, 28 and 142 nucleotides upstream from the poly(A) tail.

Thirteen other independent clones which contained A_{1a} subunit structure were also sequenced in order to investigate a heterogeneity of the A_{1a} subunit precursor proteins. The results showed that all the clones analyzed here have the completely identical nucleotide sequences to each other and encoded the glycinin precursors for the A_{1a} acidic subunit followed by the B_x basic subunit (data not shown).

Comparison of nucleotide sequences of the $A_{1a}B_x$ and the legumin A mRNAs

To compare the nucleotide sequence in the coding region of the glycinin $A_{1a}B_x$ mRNA with that of the legumin A mRNA (15) dot matrix analysis was performed. The figures for nucleotide homology did not take into account the hypothetical gaps introduced to obtain maximum homology. As shown in Fig. 3, extensive nucleotide homologies exist within both the acidic and the basic subunit components, whilst remarkable differences, which are mainly based on some relatively long deletions in the glycinin gene, occur in the nucleotide positions corresponding to the COOH-terminal region of the acidic subunit component. Figure 3 also indicated the occurrence of a nucleotide repeated unit represented as (GAX)n in both mRNA sequences. These repeated units corresponded to the consecutive glutamate or aspartate residues located in the COOH-terminal region of both acidic components. To achieve an alignment with maximum homology, a number of hypothetical gaps have been introduced into the sequences (data not shown). With this nucleotide alignment, overall, 69 % of the nucleotide positions in the $A_{1a}B_x$ mRNA was identical to the legumin A mRNA. Both nucleotide sequences corresponding to the signal peptide region show an identity in 62 % of the nucleotide positions. A comparison of the nucleotide positions in those mRNA also revealed that 67 % of the positions in the coding sequences for mature A_{1a} subunit and the corresponding region of the legumin acidic subunit, and 73 % of

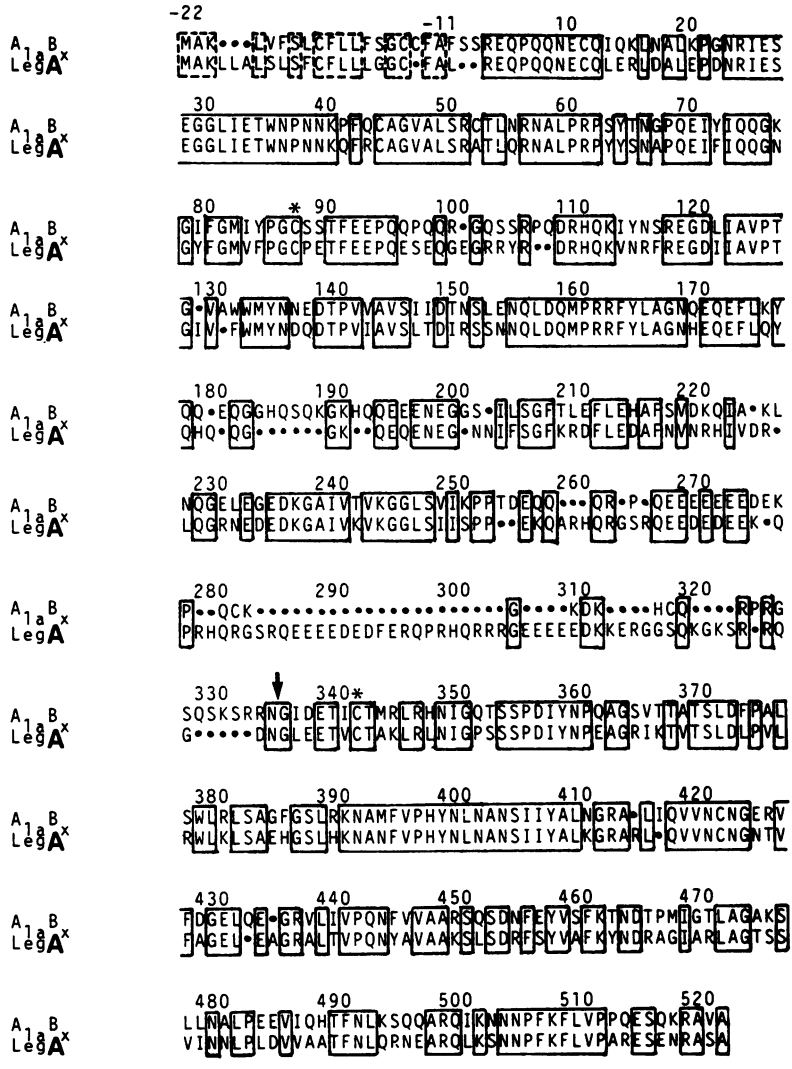

Figure 4: Comparison of the amino acid sequence of glycinin $A_{1a}B_x$ subunit precursor with that of legumin A subunit. The amino acid sequence of the legumin subunit precursor is derived from (15). The dashed boxes enclose amino acids that are identical in the signal peptide segments of the two proteins, whereas the solid boxes enclose the identical amino acids of those precursor polypeptides. The dots indicate the hypothetical deletion spaces that were placed in the each sequence to maximize the homology between the two subunit precursors. The arrow indicates the cleavage site of each protein precursor, whereas the asterisks represent the cysteine residues which form a disulfide bond in between during the posttranslational processing.

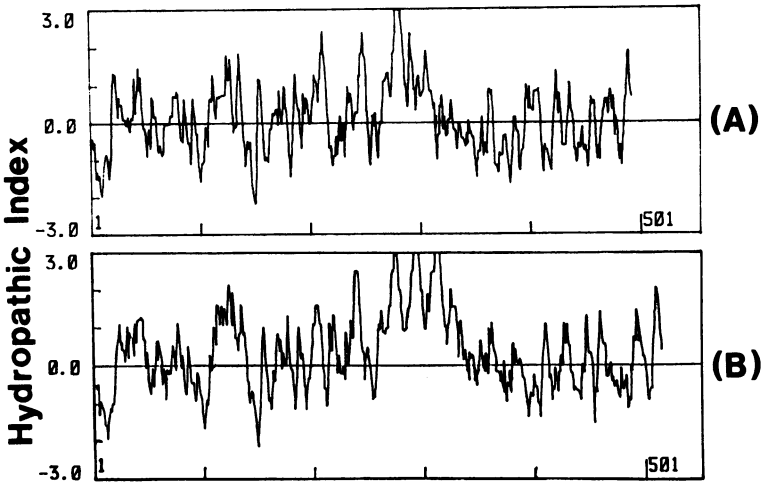

Figure 5: Comparison of the hydropathic character of glycinin $A_{1a}B_x$ subunit precursor with that of legumin A subunit precursor. The horizontal axis represents the amino acid positions in each precursor form, while the vertical axis indicates the hydropathic index. (A) $A_{1a}B_x$ subunit precursor; (B) Legumin A precursor.

those for mature B_x subunit and the corresponding region of mature basic subunit of the legumin A, were identical to each other. When compared with the nucleotide positions of the glycinin A_3 subunit family mRNAs (40 % nucleotide homology to legumin mRNAs)(6,7), the genes encoding A_2 subunit family appears to be more closely related to the legumin A gene on evolution.

Characterization of predicted protein sequence

Comparison of the amino acid sequence of glycinin $A_{1a}B_x$ subunit precursor with that of the legumin A derived from Ref. 15 is represented in Fig. 4. Using an alignment that permitted a maximum homology of amino acid sequences the overall homology in amino acid positions of these precursors is 60 %. This degree of homology is rather higher than that in the A_3 subunit family of glycinin as well as the extent of homology in nucleotide positions in between. It was also found that overall, 56 % of amino acid positions of both acidic subunit components were identical to each other, whilst 67 % of both basic subunit components were identical. The differences in amino acid sequence are not distributed uniformly throughout those protein molecules. For example, the extent of homology in the NH_2-

terminal regions of both acidic and basic subunit components is significantly higher than that in the COOH-terminal regions of them. When the entire amino acid sequences of both precursor proteins are plotted as a function of hydropathic index (16), quite similar patterns that alternated between hydrophobicity and hydrophilicity were observed as indicated in Fig. 5. On the other hand, the amino acid sequences spanning residues 259-267 of $A_{1a}B_x$ subunit precursor, and residues 251-257, 270-277, and 289-294 of the legumin A subunit precursor show a strong hydrophilicity (Fig. 5). This is one of the most outstanding feature of glycinin-like proteins. These strongly hydrophilic sequences, which consist of aspartate and glutamate, are located immediately preceding the boundary regions for the post-translational cleavage sites, and are predicted to have a predominant α-helical potential (6,7). This region is included in the repeated units found in the glycinin A_3 subunit family (6,7) and legumin subunits (15,17). Although there was no repeat of the relatively long hydrophilic sequence in the $A_{1a}B_x$ subunit precursor, the predicted secondary structure (18,19) covering the boundary region between the A_{1a} and the B_x subunit of the precursor was quite similar to those of the A_3 subunit family (6,7) and the legumin precursors (15).

DISCUSSION

In this study we analyzed fourteen independent cDNA clones covering the NH_2-terminal region of glycinin A_{1a} subunit, a member of the A_2 subunit family. The nucleotide sequences of these cDNA inserts as well as the deduced amino acid sequences were completely identical to each other. According to the classification for glycinin acidic/basic subunits (4,5), the NH_2-terminal deduced sequences of both subunits were identical to the A_{1a} and the B_{1b} subunit, respectively. However, it seems that the naming for glycinin basic subunits is not a sequence definition (31). Thus, we designate this glycinin subunit pair as $A_{1a}B_x$. When compared to the deduced protein sequence of A_2B_{1a} subunit precursor, a member of the A_2 subunit family (20), overall homology between the $A_{1a}B_x$ and the A_2B_{1a} is about 83%. On the other hand, the amino acid sequence we have deduced for $A_{1a}B_x$ is also strongly homologous to the legumin precursor

sequence deduced from the leg A gene (about 60%) (5). The $A_{1a}B_x$ is more like legumin than the glycinin A_3 subunit family (about 40%) (6.7), and legumin is less like the A_3B_4 glycinin (6) than the $A_{1a}B_x$. The termination codon (UAG) and the amino acid residue immediately following the initiator (alanine) are in glycinin A_2 subunit family and in legumin A, whereas they are UAA and glycine in the glycinin A_3 subunit family. The COOH-terminal residues of mature A_{1a} and B_x subunits were not deduced because we have no information on the nature of any linker sequence joining the acidic and the basic subunits (6.7.20). The amino acid and nucleotide sequences of glycinin and legumin subunit precursors include the following common features. 1) Most of the differences in the nucleotide sequence among those genes occur in the region including the end of the acidic subunit component, cleavage site, and the first half of the basic subunit, corresponding to the third exon. 2) The differences in this region are not distributed uniformly but occur mainly in the COOH-terminal region of the acidic subunit component. 3) Sequence divergence in the acidic component of each glycinin and legumin precursor preserves the secondary structure predicted by the Chou and Fasman algorithm (18) and the hydropathic character (16) of the amino acid sequence around the cleavage site (6,7,20). The post-translational cleavages occur on the COOH-terminal side of asparagine. 4) The positions of cysteine residues which may form disulfide bonds between the acidic and basic subunits are highly conservated, i.e. cys-85 (for A_3 subunit family) or cys-86 (for A_2 subunit family) in the acidic subunit, and cys-7 in every basic subunit component as described in (21).

Homologous regions among glycinin and legumin subunit precursors are dispersed throughout these molecules. This suggests that the similarity is not due to convergent evolution, but to divergence from a common ancestor. Recently, on the basis of structural and immunological criteria it has been speculated that the genes encoding all 11S-storage proteins found in dicots and monocots have evolved from a common ancestral gene (22-29). In order to evaluate this hypothesis it is important to accumulate structural data on the 11S proteins from a diverse group of plants. However, the result which we have obtained here

may support the above hypothesis. The extent of conservation among the 11S protein molecules may be based on intense functional constraints to evolutionary divergence, i.e. the posttranslational processing event including the intramolecular disulfide bond formation, followed by the proteolytic cleavage and effective accumulation of the proteins in protein bodies.

ACKNOWLEDGEMENTS

We are grateful to Drs. Kyuya Harada and Hisashi Hirano for their thoughtful discussions. We also appreciate excellent technical assistance of Yukie Isoyama, and manuscript preparation and editorial assistance of Michiko Murakami.

*To whom correspondence should be addressed

REFERENCES

1. Catsimpoolas, N., Kenney, J.A., Meyer, E.W. and Szuhaj, B.F. (1971) J. Sci. Food Agric. 22, 448-453.
2. Kitamura, K., Takagi, T. and Shibasaki, K. (1977) Agric. Biol. Chem. 41, 833-840.
3. Moreira, M.A., Hermodson, M.A., Larkins, B.A. and Nielsen, N.C. (1981) Arch. Biochem. Biophys, 210, 633-642.
4. Tumer, N.E., Richer, J.D. and Nielsen, N.C. (1982) J. Biol. Chem. 257, 4016-4018.
5. Staswick, P.E., Hermodson, M.A. and Nielsen, N.C. (1981) J. Biol. Chem. 256, 8752-8755.
6. Fukazawa,C., Momma, T., Hirano, H., Harada, K. and Udaka, K. (1985) J. Biol. Chem. 260, 6234-6239.
7. Momma, T., Negoro, T., Hirano, H., Matsumoto, A., Udaka, K. and Fukazawa, C.(1985) Eur. J. Biochem. 149, 491-496.
8. Marco, Y.A., Thanh, V.H., Tumer, N.E., Scallon, B.J., and Nielsen, N.C. (1984) J. Biol. Chem. 259. 13436-13441.
9. Tumer, N.E., Thanh, V.H. and Nielsen, N.C. (1981) J. Biol. Chem. 256, 8756-8760.
10. Aviv, H., and Leder, P. (1972) Proc. Natl. Acad. Sci. U.S.A. 69, 1408-1412.
11. Beaucage, S.L., and Caruthers, M.H. (1981) Tetrahedron Lett. 22, 1859-1862.
12. Birnboim. H.C., and Doly, J. (1979) Nucleic Acids Res. 5, 1513-1523.
13. Maxam, A.M., and Gilbert, W. (1980) Methods Enzymol, 65, 499-560.
14. Kreil, G. (1981) Annu. Rev. Biochem. 50, 317-348.
15. Lycett, G.W., Croy, R.R.D., Shirsat, A.H. and Boulter, D. (1984) Nucleic Acids Res. 12, 4493-4506.
16. Kyte, J. and Doolittle, R.F. (1982) J. Mol. Biol. 157, 105-132.
17. Lycett, G.W.,. Delauney, A.J., Zhao, W., Gatehouse, J.A., Croy, R.R.D. and Boulter, D. (1984) Plant Mol. Biol. 3, 91-96.
18. Chou, P.Y. and Fasman, G.D. (1974) Biochemistry 13, 223-245.

19. Staswick, P.E., Hermodson, M.A. and Nielsen, N.C. (1984) J. Biol. Chem. 259, 13424-13430.
20. Momma, T., Negoro, T., Udaka, K. and Fukazawa, C. (1985) FEBS Lett. 188, 117-122.
21. Staswick, P.E., Hermodson, M.A. and Nielsen, N.C. (1984) J. Biol. Chem. 259, 13431-13435.
22. Derbyshire, E., Wright, D.J. and Boulter, D. (1976) Phytochemistry 15, 3-24
23. Gilroy, J., Wright, D.J. and Boulter, D. (1979) Phytochemistry 18, 315-316.
24. Ohmiya, M., Hara, I. and Matsubara, H. (1980) Plant and Cell Physiol. 21, 157-167.
25. H-Nishimura, I., Nishimura, M. and Akazawa, T. (1985) Plant Physiol. 77, 747-752.
26. Yamagata, H., Sugimoto, T., Tanaka, K. and Kasai, Z. (1982) Plant Physiol. 70, 1094-1100.
27. Zhao, W-M., Gatehouse, J.A. and Boulter, D. (1983) FEBS Lett. 162, 96-102.
28. Matlashewski, G.J., Adeli, K., Altosaar, I., Shewry, P.R. and Miflin, B.J. (1982) FEBS Lett. 145, 208-212.
29. Walburg, G. and Larkins, B.A. (1983) Plant Physiol. 72, 161-165.
30. Robert, L.S., Nozzolillo, C. and Altosaar, I. (1985) Biochim. Biophys. Acta 829, 19-26.
31. Nielsen, N.C. (1983) Phil. Trans. Royal Soc. Lond. Ser. B 304, 287-296.