

# Cloning and Structural Analysis of DNA Encoding an A<sub>2</sub>B<sub>1a</sub> Subunit of Glycinin\*

(Received for publication, December 19, 1983)

Yves A. Marco, Vu H. Thanh, Nilgun E. Tumer, Bernard J. Scallan, and Niels C. Nielsen

From the United States Department of Agriculture, and the Departments of Agronomy and Biochemistry, Purdue University, West Lafayette, Indiana 47907

The partial DNA sequence of a glycinin gene in a genomic clone and a homologous cDNA clone were determined. They have nearly identical nucleotide sequences and encode the basic polypeptide and part of the acidic polypeptide for an A<sub>2</sub>B<sub>1a</sub> glycinin subunit. The protein primary structure deduced from the DNA sequence is in close agreement with the amino acid sequence of the subunit determined chemically and confirms assignment of part of the amino acid sequence in the basic component where we were able to establish an overlap using conventional approaches. The coding part of the basic subunit is interrupted by a 625-base pair A + T-rich intron whose boundaries correlate with the established consensus sequences for the exon-intron junctions. Comparison of the nucleotide sequence of the basic subunit of pea legumin gene with that of the gene for A<sub>2</sub>B<sub>1a</sub> subunit reveals 70% homology in coding regions, although there is considerably less in the 3'-flanking regions.

The preceding papers in this series (28) present the primary structure of the A<sub>2</sub>B<sub>1a</sub> subunit of glycinin. The objective of the work described in this paper was to confirm the amino acid sequence in a portion of the basic polypeptide component where overlapping peptide fragments were either not available or could not be purified. This was accomplished by determining the nucleotide sequences in DNA clones that encode this portion of the A<sub>2</sub>B<sub>1a</sub> subunit. The results of this study confirm our earlier conclusion that the acidic and basic polypeptides of a subunit are encoded by a single gene which encodes the acidic polypeptide component at its 5'-end followed by the basic component at its 3'-end (27, 31, 32). The acidic and basic components are covalently joined in the precursor by a 4-amino acid linker peptide.

## MATERIALS AND METHODS<sup>1</sup>

## RESULTS AND DISCUSSION

Two clones that contain coding sequences for an A<sub>2</sub>B<sub>1a</sub> glycinin subunit were studied. A cDNA clone, pA06, contains

a 610-bp<sup>2</sup> insert that was introduced into the *Pst*I site of pBR322. The other, a genomic clone called λAla, was selected from a genomic library that was constructed from a partial *Eco*RI digest of soybean leaf DNA. The genomic clone contains a 17-kb insert, although only a 4.1-kb *Eco*RI fragment located at one end of the insert contains coding sequence for a glycinin subunit. Details concerning the construction, selection, and characterization of these clones are described in the accompanying Miniprint Section.

The physical maps of pA06 and the 4.1-kb *Eco*RI fragment of λAla are compared in Fig. 1. With the exception of a 0.6-kb region toward the 5'-end of the 4.1-kb fragment, the restriction sites in the two inserts are identical. As data described later will show, the region in the 4.1-kb insert which does not match corresponds to an intron in the coding sequence of the genomic clone. Since Northern hybridization and hybrid selection experiments (Miniprint Section) indicate that the clones contain genetic information for glycinin, their nucleotide sequences were analyzed. The strategies employed to establish the nucleotide sequence are given in Fig. 1.

The nucleotide sequences of the coding regions in both clones are shown in Fig. 2. The two are nearly identical, although in neither case is the complete coding sequence present. Since the 5'-end of the gene is absent in the genomic clone (λAla), one cannot be certain that the information encoded within it is expressed. However, comparison of the amino acid sequence for glycinin subunit A<sub>2</sub>B<sub>1a</sub> with that deduced from the clones reveals that with two exceptions they are identical. In the case of pA06, the coding sequences extend from Thr 244 in the acidic polypeptide component until Pro 146 in the basic polypeptide component. The coding regions in genomic clone λAla extend from Asn 18 in the basic polypeptide component into the 3'-flanking region of the gene. Thus, the results confirm our previous conclusion (29, 31, 32) that both the acidic and basic polypeptide components of a glycinin subunit are part of the same transcriptional unit. The mRNA produced is initially translated into a precursor of *M<sub>r</sub>* = 60,000 which is modified post-translationally by several proteolytic cleavage steps.

Efforts to sequence the B<sub>1a</sub> polypeptide component by conventional chemical techniques (28) failed to establish an overlap between the last readable residue in cyanogen bromide fragment 3 (Gly 77) and the first residue in trypsin fragment 2 (Ala 79). However, the DNA sequence in this region clearly predicts Gly 77-Arg 78-Ala 79. Since the primary structure

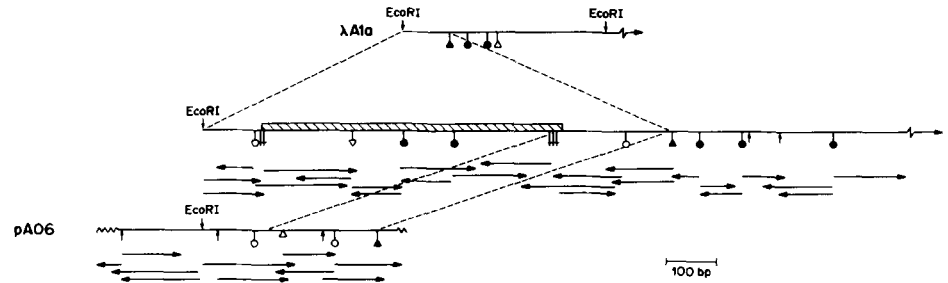
Request Document No. 83M-3573, cite the authors, and include a check or money order for \$3.60 per set of photocopies. Full size photocopies are also included in the microfilm edition of the Journal that is available from Waverly Press.

<sup>2</sup>The abbreviations used are: bp, base pairs; kb, kilobase pairs; SDS, sodium dodecyl sulfate; BSA, bovine serum albumin; PVP, polyvinylpyrrolidone; Pipes, 1,4-piperazinediethanesulfonic acid.

\* This work was supported by United States Department of Agriculture Competitive Grants Program, The American Soybean Association Foundation, cooperative research of United States Department of Agriculture-Agricultural Research Service and the Purdue Agricultural Experimental Station. Journal paper 9717. The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

<sup>1</sup> Portions of this paper (including "Materials and Methods," part of "Results," and Figs. 4-7) are presented in miniprint at the end of this paper. Miniprint is easily read with the aid of a standard magnifying glass. Full size photocopies are available from the Journal of Biological Chemistry, 9650 Rockville Pike, Bethesda, MD 20814.

FIG. 1. Restriction maps of  $\lambda$ A1a and pA06. The intron is indicated by a hatched box. Arrows indicate the strategy followed for the determination of DNA sequences. Restriction sites: *Eco*RI,  $\downarrow$ ; *Alu*I,  $\uparrow$ ; *Xor*II,  $\uparrow$ ; *Taq*I,  $\uparrow$ ; *Hinf*I,  $\uparrow$ ; *Rsa*I,  $\uparrow$ ; *San*3A,  $\uparrow$ ; *Pst*I,  $\downarrow$ .



|               |                                                                                   |      |      |      |
|---------------|-----------------------------------------------------------------------------------|------|------|------|
| pA06          | CACAGCTCCAGCCATGAGGAAGCCACAGCAAGAAGAAGATGATGATGATGAGGAAGAGCA                      | 20   | 40   | 60   |
|               | ThrAlaProAlaMetArgLysProGlnGlnGluGluAspAspAspGluGluGluGlu                         |      |      |      |
| pA06          | GCCACAGTGCCTGGAGACAGAAAGGTTGCCAACGCCAAAGCAAAAGAGCAGAAATGGCATTGATGAGACCATTTGCA     | 80   | 100  | 120  |
|               | nProGlnCysValGluThrAspLysGlyCysGlnArgGlnSerLysArgSerArgAsnGlyIleAspGluThrIleCysT  |      |      |      |
|               | Linker                                                                            |      |      |      |
| pA06          | CAATGAGACTTCGCCAAAACATTGGTCAGAATTCATCAGTACATCTACAACCTCAAGCTGGTAGCATCACAACCGCC     | 160  | 180  | 200  |
| $\lambda$ A1a | AATTCATCAGTACATCTACAACCTCAAGCTGGTAGCATCACAACCGCC                                  |      |      |      |
|               | hrMetArgLeuArgGlnAsnIleGlyGlnAsnSerSerProAspIleTyrAsnProGlnAlaGlySerIleThrThrAla  |      |      |      |
| pA06          | ACCAGCCTTGACTTCCAGCCCTCTGGCTTCTCAAACCTCAGTCCAGTATGGATCACTCCGCAAG                  | 240  | 260  | 280  |
| $\lambda$ A1a | ACCAGCCTTGACTTCCAGCCCTCTGGCTTCTCAAACCTCAGTCCAGTATGGATCACTCCGCAAGTACGTACATCATA     |      |      |      |
|               | ThrSerLeuAspPheProAlaLeuTrpLeuLeuLysLeuSerAlaGlnTyrGlySerLeuArgLys                |      |      |      |
| $\lambda$ A1a | CATGTTATAAATTAAGCTCAACAAATGTTAAAAATGGTAATAGTTATATAGTAATTAATGTTGACGACAAAGTT        | 320  | 340  | 360  |
| $\lambda$ A1a | ATCTAGCACTTGCATATGCAATTATTCATGTGCTTTTTTATATATASAAATGCCCGAAGTTATTAAGTAGTCACAGT     |      |      |      |
| $\lambda$ A1a | TTCTTGTCGATTTTACTAATCTAATTAATCTCTGCATGAGATGAATTATGCATGTGTTATTAATTTTGTACGTGT       | 400  | 420  | 440  |
| $\lambda$ A1a | GATACAATTAGTAGATAATTTGTCTTTAGTATATATGACTCATGAGTATGAAATTTTTTCAACCTTGAAAAATTTGT     |      |      |      |
| $\lambda$ A1a | TTTTGACTTTTAACCAAAAAATACCTTTGATTAATAAAAAAGAAAAATATTTTGTGCCAAAAATATTCTGATTCACAAA   | 480  | 500  | 520  |
| $\lambda$ A1a | ACATATTTTCTATGAGACTATATATGACATGTGATAAAATACCTGATAATAATTTCCATAGGATTAGGAAGAAAGGT     |      |      |      |
| $\lambda$ A1a | ATAAGATAGTGAGATAATTAATTAATATCATGATGAAAAGATAAAAGGAAAAATAATAATATTATAAAAAATTTGGATGAA | 560  | 580  | 600  |
| pA06          | TAACATGTTGTGATTAACGTACACTTGCATGGATATTGATGTATGGTCGAGAATGCTATGTTCTGGTCCCACTACACCT   | 640  | 660  | 680  |
| $\lambda$ A1a | AATGCTATGTTCTGGTCCCACTACACCT                                                      |      |      |      |
|               | AsnAlaMetPheValProHisTyrThrLe                                                     |      |      |      |
| pA06          | GAACGCGAACAGCATAATATACGCATTGAATGGGCGGGCATTGGTACAAGTGGTGAATTGCAATGGTGAGAGAGTGTG    | 720  | 740  | 760  |
| $\lambda$ A1a | GAACGCGAACAGCATAATATACGCATTGAATGGGCGGGCATTGGTACAAGTGGTGAATTGCAATGGTGAGAGAGTGTG    |      |      |      |
|               | uAsnAlaAsnSerIleIleTyrAlaLeuAsnGlyArgAlaLeuValGlnValValAsnCysAsnGlyGluArgValPheA  |      |      |      |
| pA06          | ATGGAGAGCTGCAAGAGGGAGGTGCTGATCGTTCCACAAAACCTTTCGGTGGCTGCAAAATCCAGAGCGGATAACCTT    | 800  | 820  | 840  |
| $\lambda$ A1a | ATGGAGAGCTGCAAGAGGGAGGGTGTGCTGATCGTTCCACAAAACCTTTCGGTGGCTGCAAAATCCAGAGCGGATAACCTT |      |      |      |
|               | spGlyGluLeuGlnGluGlyGlyValIleLeuValProGlnAsnPheAlaValAlaAlaLysSerGlnSerAspAsnPhe  |      |      |      |
| pA06          | GAGTACGTGTCATTCAAGACCAATGATAGACCTCGATCGGAACCTTGCAGGGGCAAACTCATTGTTGAACGCATTGCC    | 880  | 900  | 920  |
| $\lambda$ A1a | GAGTATGTGTCATTCAAGACCAATGATAGACCTCGATCGGAACCTTGCAGGGGCAAACTCATTGTTGAACGCATTGCC    |      |      |      |
|               | GluTyrValSerPheLysThrAsnAspArgProSerIleGlyAsnLeuAlaGlyAlaAsnSerLeuLeuAsnAlaLeuPr  |      |      |      |
| pA06          | A                                                                                 | 1200 | 1220 | 1240 |
| $\lambda$ A1a | AGAGGAAGTGATTGAGACACTTTTAACTAAAGAGCCAGCGGCGAGGTGAAGAACAACAACCTTTTCAGCTTCC         |      |      |      |
|               | oGluGluValIleGlnHisThrPheAsnLeuLysSerGlnGlnAlaArgGlnValLysAsnAsnProPheSerPheL     |      |      |      |
| $\lambda$ A1a | TAGTTCACCTCAGGAGTCTCAGAGGAGAGCTGTGGCTTAGAGAGCCCTTTTGCATGTGCTACCACACTTTTGTCTTT     | 1280 | 1300 | 1320 |
|               | euValProProGlnGluSerGlnArgArgAlaValAlaAMB                                         |      |      |      |
| $\lambda$ A1a | TGGGCTCTAGCTAGTGTGCTAGCAGTAATAAGAATGAATAAAAGAAACAAAGCTTAGATGCCCTTTGTTAAGTGTAG     | 1360 | 1380 | 1400 |
| $\lambda$ A1a | AATAACGGGTATAGTAACCTGTGTGTAATGAGTCACTCAAGTACCTATCGGGGAAAGCATAAATAAAGAAG           |      |      |      |
| $\lambda$ A1a | AAATCCAATAAATTTCTGTCCAATCTCTTTCTTTCTTTCTTTCTATCTTCTTGAGAACAAAATAAGGA              | 1440 | 1460 | 1480 |
| $\lambda$ A1a | AATATTGTTGATAACAATTGGTCTAAG                                                       |      |      |      |

FIG. 2. Nucleotide sequences of the pA06 and the  $\lambda$ A1a clones. The predicted amino acid sequence from both DNA sequences is indicated below the nucleotide sequences. The two sequence differences occur at positions 1043 and 1106 (\*) and do not result in an amino acid substitution.

predicted from the nucleotide sequence is in such close agreement with the protein sequence determined by Staswick *et al.* (28), we consider this assignment of residues 77–79 to be correct. Only two positions of heterogeneity between the protein and nucleic acid sequences were identified. Both of these occur in the B<sub>1a</sub> portion of the molecule. Tryptophan is replaced by Ser 43 in the sequence of B<sub>1a</sub> at a position 14 amino acids upstream from the intron. The other occurrence is the replacement of valine by Ile 164 22 amino acids before

the amber codon. The first change involves a G to C transversion in the second nucleotide of the Trp codon. The second change involves alterations in the first (G to A) and third (G to T, C or A) nucleotide of the valine codon. The variation between the sequences predicted by the clones and those determined chemically are likely related to the charge and sequence microheterogeneity discussed in more detail in the first paper of this series (28).

Comparison of the amino acid sequence determined for the

|             |                                                                                                                                                                                                               |
|-------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Pea DNA     | Lys Gly GlyLeuArgIle IleSerProProGlu LysGlnAlaArgHisGlnArgGlySerGln                                                                                                                                           |
| Soybean DNA | AAAGG TGGACTCAGGATCA TAAGCCACCCGAGA AGCAAGCGCCACAGAGAGGCAGCCAA<br>G...AAGA...G.T...G.GG...AG...G...C.GT...GTG.A.A...G.C.A...TT...<br>Glu GluAsp AspAspAspGluGluGluGln ProGlnCys ValGluThr AspLysGlyCys GI     |
| Pea DNA     | LysGlyLysSerArgArgGlnGlyAspAsnGlyLeuGluGlu ThrValCysThrAlaLysLeuArg                                                                                                                                           |
| Soybean DNA | AAAGGCAAAAGCAGAAGGCAAGGAGACAATGGGCTTGAGGAAAC AGTTTGCACTGCTAACTTCGA<br>.C..C.....A....C...CA...T..G..C. ....AATG.G.....C<br>n ArgGlnSerLysArg SerArg AsnGlyIleAspGluThr IleCysThrMetArgLeuArg<br>Linker        |
| Pea DNA     | LeuAsnIleGlyProSerSerSerProAspIleTyrAsnProGluAlaGlyArgIleLysThrValThr                                                                                                                                         |
| Soybean DNA | TTGAACATTGGCCGCTCTTCATCACCAGACATCTACAACCTGAAGCTGGTAGAATCAAACTGTTACC<br>CAA.....T.A.AA.....T.....C.....C....C...C.CC...<br>GlnAsnIleGlyGlnAsnSerSerProAspIleTyrAsnProGlnAlaGlySerIleThrThrAlaThr               |
| Pea DNA     | SerLeuAspLeuProValLeuArgTrpLeuLysLeuSerAlaGluHisGlySerLeuHisLysAsnThr                                                                                                                                         |
| Soybean DNA | AGCCTGGACCTCCAGTCTCAGGTGGCTCAAACTAAGTGCTGAGCATGGATCTCTCCACAAAATACT<br>....T...T.....CC...T..CTT.....C.....CC..T.....A...G...G...G...<br>SerLeuAspPheProAlaLeuTrpLeuLeuLysLeuSerAlaGlnTyrGlySerLeuArgLysAsnAla |
| Pea DNA     | MetPhe ValProHisTyrAsnLeuAsnAlaAsnSerIleIleTyrAlaLeuLys GlyArgAla Arg                                                                                                                                         |
| Soybean DNA | ATGTTT GTACCTCACTACAACCTGAATGCAACAGCATAATATACGCATTGAA GGGACGTGCA AGG<br>....C..G..A.....C.....C..G.....T... ..G...TT..<br>Met PheValProHisTyrThrLeuAsnAlaAsnSerIleIleTyrAlaLeuAsnGly ArgAlaLeu                |
| Pea DNA     | LeuGlnValValAsnCysAsnGlyAsnThrValPheAspGlyLysLeu GluAlaGlyArgAlaLeu                                                                                                                                           |
| Soybean DNA | CTACAGTAGTGAACGTGCAATGGCAACCGTGTGATGGAAGCTAG AAGCCGACGTGCATTG<br>.....G.....T.....TG.G.GA.....G... T.C...AG... ..GG..<br>ValGlnValValAsnCysAsnGlyGluArgValPheAspGlyGlu LeuGlnGluGly GlyVal                    |
| Pea DNA     | ThrVal ProGlnAsnTyrAlaValAlaAlaLysSerLeu AsnAspArgPheSerTyrValAla                                                                                                                                             |
| Soybean DNA | ACAGTG CCACAAACTATGCTGTGGCTGCAAGTCACTA AACGACAGTTCATATGTAGCA<br>.T.ATCGT.....T...G.....A..C. .G.G...T.AC..TGAG..T..GT..<br>LeuIleValProGlnAsnPheAlaValAlaAlaLysSer GlnSerAspAsnPheGluTyrValSer                |
| Pea DNA     | PheLysThrAsnAspArgAlaGlyIleAlaArgLeuAlaGlyThrSerSerValIleAsnAspLeuPro                                                                                                                                         |
| Soybean DNA | TTCAAGACCAATGATAGAGCTGGTATTGCAAGACTTGCAAGGACATCATAGTTATAATGATCTGCC<br>.....C.CTCG..C.G..AC.....G..AAC...T.GT.G..C.CAT...A<br>PheLysThrAsnAspArgProSerIleGlyAsnLeuAlaGlyAlaAsnSerLeuLeuAsnAlaLeuPro            |
| Pea DNA     | LeuAspValValAlaAla ThrPheLysLeuGlnArgAsp GluAlaArgGlnLeu LysSerAsn                                                                                                                                            |
| Soybean DNA | TTGGATGTGGTTGCAGCT ACATTCAAACGAGAGGATG AGGCAAGGCAGCTCA AGTCCAAC<br>GA...A...A... ..AC..T..T..C..AA....CC. .C.....G.G.AGA.CAA....<br>GluGluValIle GlnHisThrPheAsnLeuLysSer GlnGlnAlaArgGlnValLysAsnAsnAsn      |
| Pea DNA     | AsnPro PheLysPheLeuValProAlaArgGlnSerGluAsnArgAlaSerAlaAMB                                                                                                                                                    |
| Soybean DNA | AATCCCT TCAATTTCTAGTCCAGCTCGTCAGTCTGAGAACAGAGCTTCGGCTTAG ATTTCCGA<br>...T...GC..C.....C...AGG....C...GG.....GT..... AG.GCCCTTT<br>ProPheSerPheLeuValProProGlnGluSerGlnArgArgAlaValAlaAMB                      |
| Pea DNA     | CCAATCAATGAAAGTAATGAATAAGAACTAAGGCTTAGATGCCTTTGTTACTTGTGTAATAAAT                                                                                                                                              |
| Soybean DNA | TTGCATGTGCTACCACACTTTTGTCTTTTGGGCTCTAGCTAGTGTGAGCAGTAAATAAGAATGAA                                                                                                                                             |
| Pea DNA     | CGAGTCATGTACCTTTTTCGCGAAACAGAATAAAATAAGGTAAATTTTCAGTGCTC                                                                                                                                                      |
| Soybean DNA | TAAAGAAACAAAGGCTTAGATGCCCTTTGTTAAGTGTAGAATAACGGGTATAGTAA                                                                                                                                                      |

FIG. 3. Comparison of the nucleotide sequence of glycinin cDNA and genomic DNA with sequence of a cDNA encoding pea legumin (8). Pea DNA sequences are presented above the glycinin DNA sequences; only positions where nucleotide differences occur are indicated. The alignment of the two sequences are adjusted to obtain maximum homology.

mature A<sub>2</sub>B<sub>1a</sub> subunit from the seed with that for the precursor (Fig. 2) permits characterization of the linker which joins the acidic and basic components in the initial translation product. For A<sub>2</sub>B<sub>1a</sub> the linker consists of the 4 amino acids, Arg-Ser-Arg-Asn. Cleavage at the NH<sub>2</sub> terminus of the linker occurs between Lys 278, found at the COOH terminus of the mature acidic polypeptide, and arginine. Conversion of proinsulin to insulin involves hydrolysis of the same type of bond (30) and is catalyzed by a specific thiol protease recently described by Docherty *et al.* (10). Interestingly, while thiol proteases are common constituents of seeds (22), serine proteases are more prevalent in animal cells than are thiol proteases.

Explaining the mechanism for cleavage at the COOH terminus of the linker is less straightforward. An aminopeptidase could remove the linker amino acids, although this conversion is extremely specific since the NH<sub>2</sub>-terminal amino acid of every basic polypeptide studied is glycine. An alternative possibility is that a specialized protease may cleave between asparagine and glycine to form the basic component (Fig. 2). While an enzyme with this specificity is to our knowledge

unknown in higher plants, a viral protease which cleaves glutamine-glycine bonds in virus coat polypeptide precursors has been described by Kitamura *et al.* (16).

Post-translational modification also appears to occur at the COOH-terminal end of the basic polypeptide component B<sub>1a</sub>. A pentapeptide (Arg-Arg-Ala-Val-Ala) predicted from the nucleotide sequence of clone  $\lambda$ Ala is absent in the mature subunit. An interesting similarity exists between the pentapeptide and the acidic polypeptide A<sub>4</sub> in that both have a pair of NH<sub>2</sub>-terminal arginine residues. As discussed in an earlier paper (28), A<sub>4</sub> appears to be separated from A<sub>5</sub> post-translationally by a protease which cleaves just prior to the paired Arg residues. While the relationship may be coincidental, the same enzyme potentially could carry out both cleavages.

The region of the  $\lambda$ Ala clone which encodes glycinin subunit A<sub>2</sub>B<sub>1a</sub> is interrupted by a 625-bp intron (Fig. 2). The intron is characterized by long A + T regions which are atypical of the actual coding region of the gene. The intron-exon junctions at both the 5' and 3' ends have sequences equivalent to consensus sequences established for these regions in other eukaryotic genes.

Pea legumin shares many common structural features with glycinin. Both are the same size and consist of subunits with paired acidic and basic polypeptide components (8, 9). As in soybean, the initial translation products of legumin genes are precursors of about 60,000 daltons (9). Since the nucleotide sequence for a cDNA clone encoding part of a legumin subunit was recently published (8), we determined to what extent the coding sequences for these two proteins is conserved. Comparison of the two sequences (Fig. 3) reveals about 70% homology between the DNA coding sequences but only about 55% between the amino acid sequences. This is somewhat surprising since one might have anticipated that codon redundancy would have permitted more mutational drift in the nucleotide sequence than in the amino acid sequence. Most of the differences between the nucleotide sequences of the two genes involve single base substitutions, although larger substitutions (up to four bases), small deletions and insertions are also observed. While the degree of homology between the two genes is extensive in the coding region, less homology is evident when the 3'-flanking regions are compared (Fig. 2).

Structural and immunological criteria have been used in the past to establish the similarity among the 11 S proteins of legumes, and this has resulted in the notion that the genes which produce these proteins have evolved from a common ancestral gene. The high degree of homology between the legumin and glycinin genes strengthens this conclusion and argues strongly against a role for convergent evolution in the appearance of the 11 S proteins in different plant species. Recent structural data indicate that proteins similar to glycinin and legumin are found in cereals such as oats (5, 33), rice (35), and maize (7). Glycinin or legumin-like storage proteins therefore seem conserved among a rather diverse group of plants. This may indicate that the complex post-translational interconversions and packaging of these proteins into protein bodies occurs by processes which do not tolerate appreciable structural deviation. A comparison of the conserved regions among these proteins may permit a clearer definition of these fundamental structural requirements. In addition to conformational effects at the protein level, structural constraints to evolutionary divergence could be operating at an even more fundamental level. Conformational requirements in secondary structures of RNA or DNA are one means to account for the greater homology between the nucleotide sequence of legumin and glycinin coding sequences than between their amino acid sequences.

## REFERENCES

- Alwine, J. C., Kemp, D. J., and Stark, G. R. (1977) *Proc. Natl. Acad. Sci. U. S. A.* **74**, 5350-5354
- Benton, W. D., and Davis, R. W. (1977) *Science (Wash. D. C.)* **196**, 180-182
- Blattner, F. R., Williams, B. G., Blechl, A. E., Thompson, K. D., Faber, H. E., Furlong, L. A., Grunwald, D. J., Kiefer, D. O., Moore, D. D., Schumm, J. W., Sheldon, E. L., and Smithies, O. (1977) *Science (Wash. D. C.)* **196**, 161-169
- Boyer, H. W., and Roulland-Dussoix, D. (1969) *J. Mol. Biol.* **41**, 459-472
- Brinegar, A. C., and Peterson, D. M. (1982) *Plant Physiol.* **70**, 1767-1769
- Buell, G. N., Wickens, M. P., Payvar, F., and Schimke, R. T. (1978) *J. Biol. Chem.* **253**, 2471-2482
- Cross, J. W., and Adams, W. R. (1983) *J. Agric. Food Chem.* **31**, 534-538
- Croy, R. R. D., Lycett, G. W., Gatehouse, J. A., Yarwood, J. N., and Boulter, D. (1982) *Nature (Lond.)* **295**, 76-79
- Croy, R. R. D., Gatehouse, J. A., Evans, I. M., and Boulter, D. (1980) *Planta (Berl.)* **148**, 49-56
- Docherty, K., Carroll, R. J., and Steiner, D. F. (1982) *Proc. Natl. Acad. Sci. U. S. A.* **79**, 4613-4617
- Enquist, L., and Sternberg, N. (1979) *Methods Enzymol.* **68**, 281-298
- Grunstein, M., and Hogness, D. S. (1975) *Proc. Natl. Acad. Sci. U. S. A.* **72**, 3961-3965
- Heyn, R. F., Hermans, A. K., and Schilperoort, R. A. (1974) *Plant Sci. Lett.* **2**, 73-78
- Katz, L., Kingsbury, D. T., and Helinski, D. R. (1973) *J. Bacteriol.* **114**, 577-591
- Kelly, R. B., Cozzarelli, N. R., Deutscher, M. P., Lehman, I. R., and Kornberg, A. (1970) *J. Biol. Chem.* **245**, 39-45
- Kitamura, N., Semler, B., Rothberg, P., Larsen, G., Adler, C., Dornier, A., Emini, E., Hanecak, R., Lee, J., van der Werf, S., Anderson, C., and Wimmer, E. (1981) *Nature (Lond.)* **291**, 547-549
- Legerski, R. J., Hodnett, J. L., and Gray, H. B. (1978) *Nucleic Acids Res.* **5**, 1445-1464
- Maniatis, T., Hardison, R. C., Lacy, E., Lauer, J., O'Connell, C., and Quon, D. (1978) *Cell* **15**, 687-701
- Maxam, A. M., and Gilbert, W. (1980) *Methods Enzymol.* **65**, 499-560
- McGrogan, M., Spector, D. J., Goldenberg, C. J., Halbert, D., and Raskas, H. J. (1979) *Nucleic Acids Res.* **6**, 593-607
- Roychoudhury, R., Jay, E., and Wu, R. (1976) *Nucleic Acid Res.* **3**, 101-116
- Ryan, C. A., and Walker-Simmons, M. (1981) in *The Biochemistry of Plants* (Marcus, A., ed) Vol. 6, pp. 321-345, Academic Press, New York
- Sanger, F., Nicklen, S., and Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. U. S. A.* **74**, 5463-5467
- Sanger, F., and Coulson, A. R. (1976) *FEBS Lett.* **87**, 107-110
- Smith, A. J. H. (1980) *Methods Enzymol.* **65**, 560-580
- Southern, E. M. (1975) *J. Mol. Biol.* **98**, 503-517
- Staswick, P. E., Hermodson, M. A., and Nielsen, N. C. (1981) *J. Biol. Chem.* **256**, 8752-8755
- Staswick, P. E., Hermodson, M. A., and Nielsen, N. C. (1984) *J. Biol. Chem.* **259**, 13424-13430
- Staswick, P. E., Hermodson, M. A., and Nielsen, N. C. (1984) *J. Biol. Chem.* **259**, 13436-13441
- Steiner, D. F., Quinn, P. S., Chan, S. J., Marsh, J., and Tager, H. S. (1980) *N. Y. Acad. Sci.* **343**, 1-16
- Tumer, N. E., Thanh, V. H., and Nielsen, N. C. (1981) *J. Biol. Chem.* **256**, 8756-8760
- Ereken-Tumer, N., Richter, J. D., and Nielsen, N. C. (1982) *J. Biol. Chem.* **257**, 4016-4018
- Walburg, G., and Larkins, B. (1982) *Plant Physiol.* **69**, (Suppl.), 682
- Wickens, M. P., Buell, G. N., and Schimke, R. T. (1978) *J. Biol. Chem.* **253**, 2483-2495
- Yamagata, H., Sugimoto, T., Tanaka, K., and Kasai, Z. (1982) *Plant Physiol.* **70**, 1094-1100

Continued on next page.

Supplemental Material To:  
Cloning and Structural Analysis of DNA  
Encoding an A<sub>2</sub>B<sub>1a</sub> Subunit of Glycinin.

Yves A. Marco, Vu H. Thanh, Nilgun E. Tumer, Bernard J. Scallan  
and Niels C. Nielsen

# MATERIALS AND METHODS

## Materials

Terminal transferase and polynucleotide kinase were obtained from Bethesda Research Laboratories (BRL), and the restriction enzymes were purchased from either BRL or New England Biolabs. Avian myeloblastosis virus (AMV) reverse transcriptase was provided by J. Beard, Life Sciences, Inc. Nitrocellulose was obtained from Schleicher and Schuell. Gamma and alpha labelled nucleoside triphosphates were purchased from Amersham.

## Isolation of Soybean mRNA and Preparation of cDNA Clones

Polyosomal RNA was purified by two passages through an oligo (dT) cellulose column and fractionated on a sucrose gradient to purify the 18-20S mRNA as described by Tumer et al. (31). Complementary DNA copies of the 18-20S mRNA were synthesized using AMV reverse Transcriptase and an oligo (dT) primer according to the method of Buell et al. (6). Double stranded cDNA was prepared as described by Wickens et al. (34), treated with S<sub>1</sub> nuclease, and sized on a 5 to 20% sucrose gradient. The double stranded cDNA was cloned into the Pst I site of pBR322 by the oligo (dG)-oligo (dC) tailing method (21). *E. coli* K12 strain HB101 was transformed with the recombinant plasmid as described by Boyer and Roulland-Dussoix (4). Colonies with recombinant DNAs were identified by hybridization to <sup>32</sup>P-labelled cDNA according to Grunstein and Hogness (12). Plasmid DNA was purified according to the method of Katz (14).

## Identification of cDNA and Genomic Clones by Hybrid Selection

50 µg recombinant DNA from cDNA or phage was digested with Eco RI, denatured with NaOH, and blotted onto nitrocellulose (20). The nitrocellulose filters were hybridized at 47°C for 2 hours with 10 µg of poly A<sup>+</sup> RNA in 300 µl in a solution containing 30% formamide, 50 mM Pipes, pH 7.6, 500 mM NaCl, 2 mM EDTA, 0.4% SDS, 100 µg/ml poly A<sup>+</sup> RNA and 200 µg/ml *E. coli* tRNA. After hybridization, the filters were washed 3 times with 5 ml of 10 mM Tris-HCl, pH 7.5, 2 mM EDTA at 4°C for 5 minutes, and then were washed at 45°C for 3 minutes with the elution buffer containing 2 mM EDTA, pH 7.5, and 10 µg/ml *E. coli* tRNA. Elution of the mRNA bound to the filters was carried out with 600 µl of the elution buffer at 100°C for 2 minutes. The mRNA was translated in a reticulocyte lysate translation system and analyzed on a SDS-polyacrylamide gel as described by Tumer et al. (31).

## Identification of cDNA and Genomic Clones by Northern Blotting

Poly A<sup>+</sup> RNA was fractionated on 1% agarose containing 5 mM methylmercury and blotted onto a nitrocellulose filter according to Alwine et al. (11). DNA from cDNA and genomic clones was labelled by nick translation (15). Sections of the nitrocellulose were hybridized for 14 hr. at 42°C with <sup>32</sup>P-labelled DNA in 50% formamide 5 x SSC, 50 mM potassium phosphate, pH 6.5, 0.1% JSA, 0.1% Ficoll, 0.1% PVP, 250 µg/ml salmon sperm DNA, 0.1% SDS and 10 µg/ml polyadenylic acid. Following hybridization, the filters were washed with 2 x SSC, 0.1% SDS and exposed to <sup>32</sup>P. Southern blotting was performed after separating the DNA fragments on agarose, transferring to nitrocellulose (26) and hybridizing as described above.

## Preparation of the Soybean Genomic Library

DNA was isolated from soybean (C1635-1-1-1) leaves as described by Heyn et al. (13). The library was constructed from a partial Eco RI digest of soybean DNA as described by Maniatis et al. (18). Packaging of the recombinant phage was accomplished by a modification of the method of Enquist and Sternberg (11, 5, Gupta, personal communication). Screening of the library was performed by the method of Benton and Davis (2). Phage were grown in NZYM medium (5g NaCl, 2g MgCl<sub>2</sub>, 10g NZ-amine and 5g yeast extract per liter of water) and purified according to Blattner et al. (3). Phage DNA was extracted by the procedure of Maniatis (18).

## DNA Sequencing

The DNA sequences were either determined according to the method of Maxam and Gilbert (19), or the Sanger technique was used with mp7 as a vector (23-25).

## Physical Mapping of $\lambda$ Ala

The Eco RI fragments of  $\lambda$ Ala were mapped with Bal-31 exonuclease. Treatment of DNA with exonuclease and generation of Eco RI fragments were done as described by Legerki et al. (17). The 5'-terminal of the fragments were labelled with <sup>32</sup>P-ATP, separated by electrophoresis in 0.7% agarose, blotted onto nitrocellulose (26) and visualized by autoradiography. This permitted the loss of Eco RI fragment to be followed as a function of digestion time.

## RESULTS

### Synthesis and Isolation of cDNA and Genomic Clones

We have previously demonstrated that total poly A<sup>+</sup> RNA from midmaturation stage embryos contains two superabundant mRNA classes of about 18S and 20S which encode glycinin and  $\beta$ -conglycinin subunits, respectively (31). A fraction enriched for these messages was purified by a sucrose gradient and used as template to synthesize cDNA. Double stranded cDNAs were prepared according to standard procedures (6), treated with S<sub>1</sub> nuclease and sized on 5 to 20% linear sucrose gradients. The cDNA fraction larger than 1.3 kb was collected and inserted into the Pst I site of pBR322 by the oligo (dG)-oligo (dC) tailing method (21). *Escherichia coli* K12 strain HB101 was transformed with the recombinant plasmids, and the transformants were screened for tetracycline resistance and ampicillin sensitivity. Four clones pA06, pA36, pA26 and pA40 hybridized strongly to <sup>32</sup>P-labelled cDNA synthesized from 18-20S mRNA. Plasmid pA06, which exhibited the strongest hybridization, was selected for further study. Pst I cleavage and subsequent gel electrophoresis of pA06 showed that it contained a 610 bp insert. A restriction map of the insert in pA06 was subsequently determined by conventional methods (Fig. 1).

A genomic library was constructed in  $\lambda$ Charon 4A using soybean DNA partially cleaved with Eco RI. It was screened with <sup>32</sup>P-labelled cDNA prepared from 18-20S enriched poly A<sup>+</sup> RNA. A clone,  $\lambda$ Ala, was identified by plaque hybridization and mapped both by restriction site analysis and the Bal-31 exonuclease technique of Legerki et al. (17). The clone contained a 17 kb insert which produced 5 fragments after Eco RI digestion. These fragments were about 11.5, 4.1, 0.7, 0.55 and 0.1 kb in size. The 0.7 kb fragment was the one most insensitive to Bal-31 exonuclease, followed by the 11.5 kb fragment. This indicated that the 0.7 kb fragment was probably adjacent to the long arm of the Charon 4A cloning vehicle while the 11.5 kb fragment occupied the central part of the insert (Fig. 4). The 4.1 and 0.55 kb fragments were equally sensitive to Bal-31, but were degraded more rapidly than the 11.5 and 0.7 kb fragments. This result indicated that these two fragments were adjacent to the short arm of the cloning vehicle.

The orientation of the 0.55 and 4.1 kb fragments was determined by characterizing restriction fragments which crossed the junction between the genomic DNA and short arm of the cloning vehicle (Fig. 4). Digestion of  $\lambda$ Ala with Bgl II produced three restriction fragments (e.g., 4.5, 4.7 and about 9 kb) not present in digests of the Charon 4A arms. Each was purified and further degraded with Eco RI. In the case of the 4.5 kb fragment, this treatment produced 2.3, 0.55 and 1.6 kb pieces of DNA. The 1.6 kb piece originated from the short arm (3) while the 2.3 kb fragment came from the 4 kb Eco RI fragment of the genomic insert in  $\lambda$ Ala (cf. Figs. 1 and 2). Not only was the 2.3 kb fragment able to hybridize with pA06 in southern blots, but its position in the fragment was established by nucleotide sequence analysis (Fig. 2). The 0.55 kb fragment is therefore located between the 4 kb Eco RI fragment of the genomic DNA and the short arm of Charon 4A. Eco RI digestion of the 4.7 kb Bgl II fragment of  $\lambda$ Ala produced 2 pieces of DNA of 2.9 and 1.8 kb. The 1.8 kb piece originated from the 4.1 kb Eco RI fragment of genomic DNA in  $\lambda$ Ala. The 2.9 kb piece therefore extended into the 11.5 kb Eco RI fragment. The third unique Bgl II fragment of  $\lambda$ Ala (e.g., 9 kb) decreased only slightly in size when digested with Eco RI. This is consistent with the conclusion that the 0.7 kb Eco RI fragment followed the 11.5 kb Eco RI fragments. Experiments with the purified 0.7 kb fragment revealed that it was cleaved by Bgl II.

The Xor II restriction fragment across the junction between the insert and short arm was also studied. In performing this experiment use was made of the fact that the 4.1 kb Eco RI fragment in  $\lambda$ Ala contained a single Xor II site which was located after the large intron which interrupted the coding sequence for the basic polypeptide component (Figs. 1 and 2). When pA06 was used as a probe in southern blots, it hybridized with the 900 bp segment of DNA between the Xor II site and the Eco RI site at the 5'-end of the A<sub>2</sub>B<sub>1a</sub> coding sequence in  $\lambda$ Ala. Digestion of  $\lambda$ Ala with Xor II resulted in a 3.2 kb fragment which by southern analysis hybridized to the pA06 probe. Purification of the 3.2 kb Xor II fragment and its subsequent digestion with Eco RI resulted in formation of the 0.37 kb piece of DNA from the 4.1 kb Eco RI fragment, the 0.55 kb Eco-Eco fragment and a 1.7 kb piece of DNA from the short arm of the Charon 4A short arm (3). The results of this experiment therefore confirmed the assignment of the 0.55 kb Eco RI fragment as being located next to the short arm of Charon 4A.

Only the 4.1 kb Eco RI fragment hybridized with <sup>32</sup>P-labelled cDNA probes synthesized from 18-20S enriched polyadenylated mRNA (Fig. 5). This suggested that it was the only one which contained coding sequence for one of the prevalent seed proteins. To further characterize the 4.1 kb Eco RI fragment, it was subcloned into pBR322. This fragment and the cDNA insert from pA06 were labelled by nick translation and hybridized to poly A<sup>+</sup> RNA from developing seeds (Fig. 6). The northern analysis showed that both the cDNA insert and the 4.1 kb genomic DNA fragment hybridized to 18S poly A<sup>+</sup> RNA. This class of prevalent seed messages was previously shown to code for glycinin precursors (31). Northern hybridization analysis of the other cDNA clones indicated that pA40 also hybridized to 18S mRNA, pA26 hybridized to 16S mRNA while pA36 hybridized to 20S mRNA. The 20S messages were previously shown to encode  $\beta$ -conglycinin precursors (31).

Hybrid-select translation was used to confirm that pA06 and  $\lambda$ Ala coded for glycinin precursors. The recombinant plasmid from pA06 and DNA from  $\lambda$ Ala were linearized with Eco RI, bound to nitrocellulose and then hybridized to soybean poly A<sup>+</sup> RNAs from midmaturation stage embryos. The specifically bound mRNAs were eluted by increasing the temperature to 100°C and these were translated *in vitro*. The results are indicated in Figure 7. The translation products of enriched soybean poly(A)RNA are in the second lane. A group of proteins around 60,000 molecular weight previously identified as glycinin precursors were major translation products (32). The  $\lambda$ Ala DNA selected mRNA that for the most part directed the synthesis of one of these precursors (lane 3), and it was specifically precipitated by antibodies against glycinin (lane 4). The clone pA06 also selected mRNA which caused a 60,000 molecular weight precursor to be synthesized. These results suggested that both pA06 and 4.1 kb fragment from  $\lambda$ Ala encoded part of a glycinin subunit.

As summarized in the main body of the manuscript, nucleotide sequence analysis indicated that coding sequence for glycinin subunit A<sub>2</sub>B<sub>1a</sub> began at the border of the 4.1 kb Eco RI fragment closest to the short arm of Charon 4A and continued 3' into that fragment (Fig. 2). The 0.55 kb fragment which lay between the 4.1 kb Eco RI fragment and the short arm neither hybridized with cDNA prepared from glycinin enriched mRNAs (Fig. 5) nor contained coding sequence for the NH<sub>2</sub>-terminal end of the A<sub>2</sub>B<sub>1a</sub> precursor (data not shown). The insert in  $\lambda$ Ala could be part of a pseudogene for a glycinin subunit. However, since fragments from a partial Eco RI digest of soybean DNA were used to prepare the library, it is considered more likely that the 0.55 kb fragment became situated next to the 4 kb fragment as a cloning artifact. Thus, while clones pA06 and  $\lambda$ Ala were quite useful in our efforts to determine the primary structure of the A<sub>2</sub>B<sub>1a</sub> subunit by conventional approaches to protein sequence analysis, other clones with complete coding sequences will have to be characterized in their entirety to define unambiguously the features of these genes.

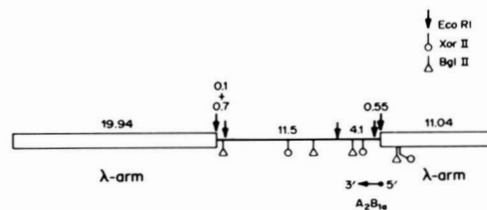


Figure 4. Partial restriction map of  $\lambda$ Ala genomic clone. The Eco RI fragments were mapped by restriction analysis, as well as by the Bal-31 exonuclease method of Legerki et al. (17). The Bgl II and Xor II sites were determined by double digestion methods. Figure 1 gives a more detailed map of the 4.1 kb Eco RI fragment which was subcloned into pBR322.

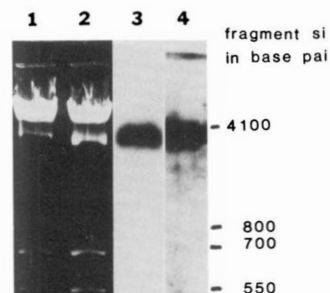


Figure 5. Hybridization of nick translated cDNA with Eco RI digested phage  $\lambda$ Ala DNA. Phage DNA was digested with Eco RI, transferred to nitrocellulose and hybridized to nick translated cDNA (sp. act.  $2 \times 10^6$  cpm/µg). A 9 kb genomic DNA fragment comigrating with the phage arms is not visible on this gel: lane 1, electrophoretic separation of partially Eco RI digested  $\lambda$ Ala DNA; lane 2, electrophoretic separation of total Eco RI digest; lane 3, autoradiograph of lane 1 after hybridization to nick translated cDNA; and lane 4, autoradiograph of lane 2 after hybridization to nick translated cDNA. The 800 bp fragment in lane 1 probably is a partial digestion product that originated from the left border of the insert (e.g., 0.7 & 0.1 kb fragments, see Fig. 4).

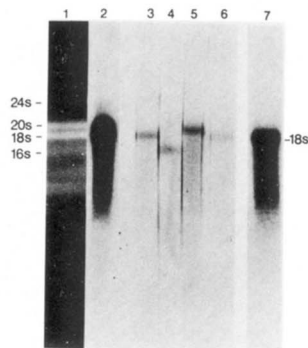


Figure 6. Hybridization of cDNA clones and 4.1 kb Eco RI fragment with total poly(A)RNA. Poly A RNA was fractionated on a 1% agarose gel that contained 5 mM methyl mercury, transferred onto nitrocellulose filter and hybridized to nick translated pA06, pB36, pC26 and B40 DNAs (sp. act. 10<sup>7</sup> cpm/μg) and 4.1 kb Eco RI fragment (sp. act. 10<sup>7</sup> cpm/μg): lane 1, ethidium bromide stained gel showing electrophoretic separation of poly A RNA; lane 2, autoradiogram of cDNA synthesized from the mRNA shown in lane 1; lane 3, autoradiogram of Northern blot hybridized to pA06 DNA; lane 4, autoradiogram of Northern blot hybridized to pC26 DNA; lane 5, autoradiogram of Northern blot hybridized to pB36 DNA; lane 6, autoradiogram of Northern blot hybridized to pB40 DNA; and lane 7, autoradiogram of Northern blot hybridized to 4.1 kb Eco RI fragment.

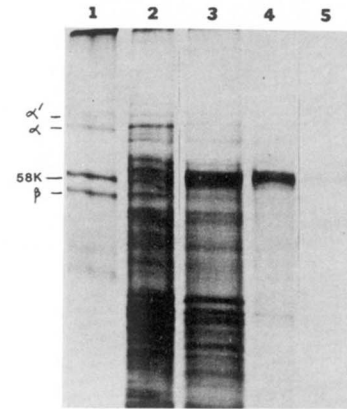


Figure 7. *In vitro* translation of hybrid-selected mRNA. The cDNA clone, pA06, or the 4.1 kb genomic Eco RI fragments from JA1a were bound to nitrocellulose and hybridized to a mixture of poly A RNA. Specifically bound mRNA was eluted by increasing the temperature to 100°C, precipitated and translated in a rabbit reticulocyte lysate translation system as described in Materials and Methods: lane 1, standard soybean proteins showing α, α' and β subunits of 75 and 58K glycinin subunit complexes; lane 2, translation products of total poly A RNA; lane 3, translation products from mRNA selected by the 4.1 kb Eco RI fragment; lane 4, immunoprecipitation of lane 3 by glycinin antibodies; and lane 5, translation of mRNA selected by pA06 DNA.