

Characterization of the Glycinin Gene Family in Soybean

Niels C. Nielsen,^{a,1} Craig D. Dickinson,^{a,2} Tae-Ju Cho,^{a,3} Vu H. Thanh,^{a,4} Bernard J. Scallan,^{a,5} Robert L. Fischer,^{b,6} Thomas L. Sims,^{b,7} Gary N. Drews,^b and Robert B. Goldberg^{b,1}

^aU.S. Department of Agriculture/Agricultural Research Service, Department of Agronomy, Purdue University, West Lafayette, Indiana 47907

^bDepartment of Biology, University of California, Los Angeles, California 90024-1606

We characterized the structure, organization, and expression of genes that encode the soybean glycinins, a family of storage proteins synthesized exclusively in seeds during embryogenesis. Five genes encode the predominant glycinin subunits found in soybeans, and they have each been cloned, sequenced, and compared. The five genes have diverged into two subfamilies that are designated as Group-I and Group-II glycinin genes. Each glycinin gene contains four exons and three introns like genes that encode related proteins in other legumes. Two other genes have been identified and designated as “glycinin-related” because they hybridize weakly with the five glycinin genes. Although not yet characterized, glycinin-related genes could encode other glycinin subunit families whose members accumulate in minor amounts in seeds. The three Group-I glycinin genes are organized into two chromosomal domains, each about 45 kilobase pairs in length. The two domains have a high degree of homoeology, and contain at least five genes each that are expressed either in embryos or in mature plant leaves. Gel blot studies with embryo mRNA, as well as transcription studies with ³²P-RNA synthesized in vitro from purified embryo nuclei, indicate that glycinin and glycinin-related genes become transcriptionally activated in a coordinated fashion early in embryogenesis, and are repressed coordinately late in seed development. In addition to transcriptional control processes, posttranscriptional events also are involved in regulating glycinin and glycinin-related mRNA levels during embryogenesis.

INTRODUCTION

Glycinins are the predominant storage proteins in soybean seeds. They account for more than 20% of the seed dry weight in some cultivars, have no known catalytic activity, and are thought to function as a reserve for carbon and nitrogen to be used upon seed germination (Spencer, 1984). Glycinin proteins are produced primarily in cotyledon cells where they are sequestered within subcellular organelles called protein bodies. The structures of glycinin proteins have been studied extensively. As isolated from seed extracts, the glycinins are an oligomer of six similar subunits (Badley et al., 1975). The properties of these subunits have been reviewed extensively (Millerd, 1975;

Derbyshire et al., 1976; Wolf, 1976; Larkins, 1981; Nielsen, 1984), and five major subunits have been identified on the basis of differences in their primary structures (Moreira et al., 1979). Additional glycinin subunits may also be present at low levels in dry seeds (Lei et al., 1983). Each glycinin subunit is composed of two disulfide-linked polypeptides. One polypeptide has an acidic isoelectric point, and the other is basic. The two polypeptide chains result from posttranslational cleavage of proglycinin precursors (Turner et al., 1982), a step that occurs after the precursor enters the protein bodies (Chrispeels et al., 1982).

Glycinin subunits accumulate rapidly during embryogenesis, and this accumulation is associated with dramatic changes in the prevalence of glycinin mRNAs. Glycinin mRNAs begin to accumulate early in embryogenesis, are highly prevalent during the midmaturation stage, and then decay prior to seed dormancy (Goldberg et al., 1981a, 1981b; Meinke et al., 1981). The accumulation and decay of these mRNAs is regulated in part by transcriptional processes similar to those that regulate other seed protein mRNAs (Goldberg et al., 1981a; Walling et al., 1986).

Several glycinin genes have been cloned and studied. Fischer and Goldberg (1982) described three highly homologous, nonallelic, glycinin genes designated as Gy₁, Gy₂, and Gy₃, and Scallan et al. (1985) reported the exist-

¹ To whom correspondence should be addressed.

² Current address: Department of Biology, University of California, La Jolla, CA 92093.

³ Current address: Department of Biochemistry, Chung-Buk National University, Cheong-ju, Korea.

⁴ Current address: Department of Biological Science, Stanford University, Stanford, CA 94305.

⁵ Current address: Department of Molecular Genetics, Hoffman-LaRoche, Inc., Nutley, NJ 07110.

⁶ Current address: Division of Molecular Plant Biology, 313 Hilgard Hall, University of California, Berkeley, CA 94720.

⁷ Current address: Department of Botany, Ohio State University, Columbus, OH 43210.

ence of two additional genes that were named *Gy₄* and *Gy₅*. Because the glycinin subunits constitute a large proportion of the total seed protein, they are of considerable nutritional and economic importance to man and are an important target of efforts to improve seed quality. Success in these endeavors depends upon a detailed understanding of glycinin gene structure and the mechanisms that regulate glycinin gene expression. In this paper we show that glycinin genes have similar structures and that their expression is coordinately regulated by transcriptional processes during seed development. We further show that the *Gy₁*, *Gy₂*, and *Gy₃* glycinin genes are organized into two related chromosomal domains, and that these domains also contain several unrelated genes that are expressed in mature plant leaves.

RESULTS

Several Glycinin and Glycinin-Related Genes Are Present in the Soybean Genome

Table 1 shows the nomenclature used to describe the five glycinin genes (*Gy₁* to *Gy₅*) and the subunits they encode. Table 1 also specifies the group or subfamily designation for each gene (Nielsen, 1984) and provides a guide to identify glycinin subunits that parallels gene designations (e.g. *Gy₁* encodes a G1 subunit). The glycinin genes *Gy₁*, *Gy₂*, and *Gy₃* originated from an *AluI*/*HaeIII* genomic DNA library from the variety Dare. They were identified using a glycinin cDNA clone (A-28) as a probe (Goldberg et al., 1981a; Fischer and Goldberg, 1982). The isolation of a genomic clone containing *Gy₄* (also from Dare) was described by Scallan et al. (1985), and was identified using cDNA clone pG258 as a probe. The *Gy₅* glycinin gene originated from a genomic library of genomic DNA from the variety Forrest, and was identified using the cDNA clone pG23 as a probe (Scallan et al., 1985). These five genes encode the glycinin subunits that were identified on the basis of differences in their amino acid sequences (Moreira et al., 1979, 1981; Staswick et al., 1981, 1984).

Table 1. Glycinin Gene Family

Gene	Group	Subunit	<i>M_r</i> ^a	Molecular Weight ^b
<i>Gy₁</i>	1	G1 (A ₁ B ₂)	58,000	55,700
<i>Gy₂</i>	1	G2 (A ₂ B _{1a})	58,000	54,400
<i>Gy₃</i>	1	G3 (A _{1b} B _{1b})	58,000	54,300
<i>Gy₄</i>	2	G4 (A ₅ A ₄ B ₃)	69,000	63,700
<i>Gy₅</i>	2	G5 (A ₃ B ₄)	62,000	58,000

^a Calculated as the sum of acidic and basic polypeptides as determined by SDS electrophoresis.

^b Calculated from derived amino acid sequences.

The calculated and experimentally determined molecular weights of the glycinin subunits are also shown in Table 1, as is the nomenclature used previously to specify acidic and basic polypeptide chains purified from each subunit.

We identified cDNA clone pG3-1 by a low stringency (30°C) screen of a midmaturation stage embryo cDNA library (Goldberg et al., 1981a) with a mixed *Gy₁*, *Gy₂*, and *Gy₃* probe. Phage DNA blots showed that pG3-1 reacted with the *Gy₁*, *Gy₂*, and *Gy₃* genes, but was most homologous to the *Gy₃* gene (G.N. Drews and R.B. Goldberg, unpublished data). Genomic DNA blots indicated that the pG3-1 cDNA clone hybridized weakly at 42°C to *Gy₃* glycinin restriction fragments, as well as to other genomic DNA fragments that did not react with either Glycinin-I or Glycinin-II glycinin gene probes (data not shown). RNA blot analysis showed that pG3-1 hybridized with a 2.1-kb embryo mRNA (data not shown) whose concentration increased and decreased during embryogenesis in a manner resembling glycinin mRNAs. Because the pG3-1 cDNA clone hybridized with glycinin gene sequences and a glycinin-sized mRNA, we designated it as a glycinin-related cDNA clone that could represent an additional glycinin gene sequence.

The Group-I Glycinin Genes Are Organized into Two Chromosomal Domains

To study the organization of Group-I glycinin genes, a library of leaf nuclear DNA fragments in the Charon 4 vector was screened by plaque hybridization using glycinin cDNA clone A-28 (see "Methods"). Fifty-two genomic clones were recovered from the screen and were represented by the 6 overlapping phages λDA28-1, λDA28-30, λDA28-4, λDA28-25, λDA28-6, and λDA28-26. Figure 1 shows that these phages define two distinct domains in the soybean genome. One domain contained two linked glycinin genes (*Gy₁* and *Gy₂*) and the other contained a single glycinin gene (*Gy₃*). We expanded these domains by rescreening the Dare genomic library with recloned single copy boundary restriction fragments. The 48 clones recovered from the walking experiment were represented by the phages λDA-3, λDC-3, λDD-6, and λDD-7 (Figure 1). Figure 1 summarizes the restriction maps, location, and transcriptional polarity of the glycinin genes contained within these domains. Together these chromosomal regions represented approximately 85 kb of the soybean genome.

We investigated phages that spanned both domains (Figure 1) to determine whether they contained other seed protein and nonseed protein genes (Fischer and Goldberg, 1982). Clones representing the soybean β-conglycinin, lectin, and Kunitz trypsin inhibitor genes did not hybridize with any of the phages, indicating that members of these seed protein gene families were not tightly linked to the three Group-I glycinin genes. By contrast, the pG258 and pG23 Group-II glycinin cDNA clones hybridized weakly

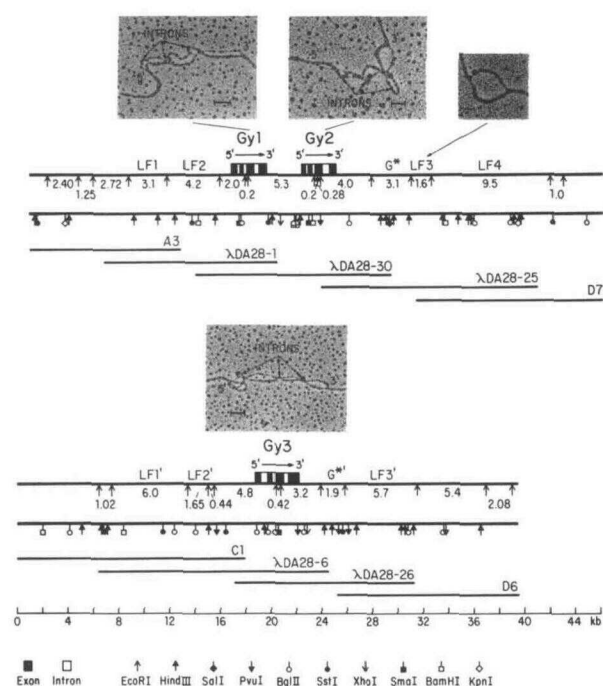


Figure 1. A Molecular Map of Two Soybean Chromosomal Domains.

The map is defined by 9 overlapping genomic clones and contains three glycinin genes (Gy_1 , Gy_2 , and Gy_3). Recombinant phages λ DA28-1 (13.6-kb insert), λ DA28-30 (15.4 kb), λ DA28-25 (18 kb), λ DA28-6 (17.9 kb), and λ DA28-26 (14.0 kb) were isolated by screening 500,000 plaques of a soybean *AluI*/*HaeIII* Charon 4 library using cDNA clone A-28 as probe (Fischer and Goldberg, 1982). Clones λ DA-3 (11.3 kb), λ DC-1 (17.7 kb), λ DD-6 (14.2 kb), and λ DD-7 (14.7 kb) were isolated by screening the soybean library (150,000 plaques) with recombined restriction fragments from the previously isolated phages. The location of the glycinin genes was determined by DNA gel blot studies or by DNA sequencing and S1 protection experiments (data not shown). Transcriptional polarities were determined by comparing DNA gel blot studies with random or oligo(dT)-primed hybridization probes (Fryberg et al., 1980), and was confirmed by DNA sequence analysis (Figure 4). The exon (black) and intron (white) regions were determined by analysis of R-loops and confirmed by DNA sequencing. R-loops were generated by reacting an excess of midmaturation embryo mRNA (Goldberg et al., 1981a, 1981b) to phage DNAs λ DA28-1, λ DA28-4, and λ DA28-26 phage DNAs. The sizes of the introns and exons were similar to those indicated in Figure 4. Gy_1 : exon 1 (E1), 380 ± 100 bp; intron (I1), 200 ± 55 bp; E2, 260 ± 60 ; I2, 240 ± 60 ; E3, 420 ± 90 ; I3, 420 ± 90 ; E4, 670 ± 140 ; ($n = 25$). Gy_2 : E1, 400 ± 100 ; I1, 300 ± 90 ; E2, 290 ± 80 ; I2, 350 ± 100 ; E3, 690 ± 150 ; I3, 780 ± 180 ; E4, 580 ± 140 ($n = 2$). Gy_3 : E1, 410 ± 100 ; I1, 590 ± 150 ; E2, 270 ± 80 ; I2, 300 ± 60 ; E3, 620 ± 160 ; I3, 460 ± 100 ; E4, 640 ± 140 ($n = 18$). G^* identifies restriction fragments that are glycinin-related and hybridize weakly to glycinin cDNA clones pG23 and pG258 representing Group-II glycinin genes (Scallan et al., 1985). R-loops of the LF3 gene in λ FA28-5 (Fischer and Goldberg, 1982) and embryo mRNA were formed as described above, and were 780 ± 160 bp in size ($n = 17$). The bar sizes in micrographs of the R-loops equal 200 bp.

with a DNA region immediately downstream from both the Gy_2 and Gy_3 glycinin genes (Figure 1). DNA fragments in this region did not detectably react at low stringency (22°C) with Group-I glycinin cDNA probes or with the pG3-1 glycinin cDNA clone (data not shown). By contrast, these fragments reacted with a 2.1 glycinin-sized mRNA that was stable at a high wash temperature (72°C). Together, these data indicate that the Gy_1 , Gy_2 , and Gy_3 glycinin genes are not linked to other seed protein genes, but are linked to divergent glycinin sequences related distantly to the Group-II subfamily. We designated these genomic sequences as G^* , or glycinin-related sequences, to distinguish them from Group-I and Group-II glycinin genes.

Hybridization of leaf cDNA (Fischer and Goldberg, 1982) with phages representing the glycinin gene domains indicated that genes expressed in the mature plant leaf were also present. As shown in Figure 2A, *EcoRI* restriction fragments designated as LF1, LF1', LF2, LF2', LF3, LF3', LF4, and LF4' (Figure 1) that flank the glycinin genes hybridized specifically with labeled leaf cDNA. We previously showed that the LF3 gene encodes 0.9-kb mRNA that represents $7 \times 10^{-4}\%$ (4 molecules/cell) of the leaf mRNA mass and is not present in embryo mRNA (Fischer and Goldberg, 1982). Figure 2B shows this result. We also reacted an LF4 gene probe with a leaf mRNA gel blot and determined that it hybridized with a 2.5-kb mRNA that constituted $2.5 \times 10^{-4}\%$ (0.5 molecules/cell) of the leaf mRNA mass (R.L. Fischer and R.B. Goldberg, unpublished results). Because the LF3 and LF4 restriction fragments generated hybridization signals with leaf cDNA that were similar in intensity to those observed with the other DNA fragments (Figure 2A), all of the linked leaf genes probably encoded rare-class mRNAs. Together, these results indicate that the glycinin genes are closely linked to genes that are not expressed during embryogenesis but are expressed in leaves of the mature plant.

The Glycinin Gene Domains Are Homoeologous

Figure 1 shows that the *EcoRI* restriction fragments (e.g. LF1/LF1', LF2/LF2') that hybridize with the leaf cDNA are located in similar positions relative to the glycinin genes in the two domains. To determine whether these fragments contained analogous genes, we measured the sizes of mRNAs encoded by the LF3 and LF3' regions. Figure 2B shows that both LF3 and LF3' hybridized with 0.9-kb leaf mRNAs. This result suggested that the two glycinin gene domains contained extensive regions of homology. To test this conclusion directly, we formed heteroduplexes between representative cloned phage DNAs that contained each domain and visualized these duplexes in the electron microscope. The phages used in these experiments are summarized in Figure 3A. Figure 3, B to F, shows that regions of homology (double-stranded) and nonhomology (single-stranded) are interspersed throughout the 42 kb contained in each of the two domains. Even at the bound-

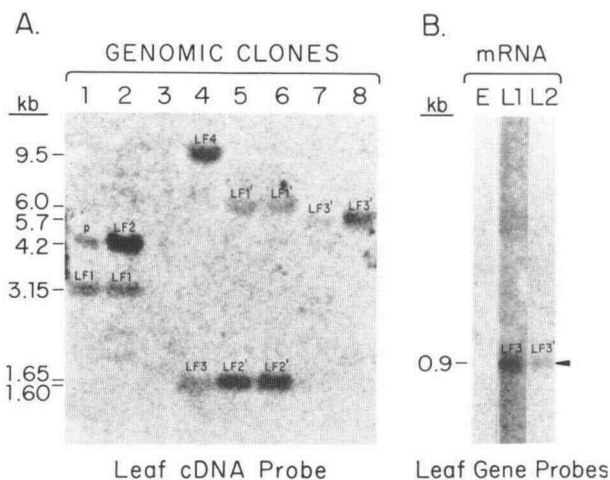


Figure 2. Leaf Genes That Flank Cloned Glycinin Genes.

(A) Hybridization of genomic clones with labeled leaf cDNA. Phage DNA (2.5 μ g) was digested with EcoRI, electrophoresed in 0.75% agarose, blotted, and hybridized (1 M Na⁺, 50% formamide, 42°C) with random-primed leaf ³²P-cDNA. LF1, LF2, LF3, LF4, LF1', LF2', LF3', and LF4' refer to restriction fragments that hybridized with the cDNA probe (see Figure 1). Incomplete digestion of the DNA with EcoRI results in the band denoted "p". Lane 1, λ DA-3; lane 2, λ DA28-1; lane 3, λ DA28-30; lane 4, λ DA28-25; lane 5, λ DC-1; lane 6, λ DA28-6; lane 7, λ DA28-8; lane 8, λ DD-6. The λ DA28-8 phage overlaps phages λ DA28-30, λ DA28-25, and λ DD-7.

(B) Hybridization of cloned LF3 and LF3' sequences to leaf and embryo mRNA. The mRNAs were separated by electrophoresis in gels that contained methylmercury hydroxide, transferred to nitrocellulose, and hybridized to labeled DNA. Hybridization and wash conditions are described in "Methods". Leaf mRNA sizes were estimated relative to rRNA and embryo superprevalent mRNA size standards (Goldberg et al., 1981a, 1981b). Lane E, embryo mRNA hybridized to labeled pG26R5.4 (recloned 5.4-kb EcoRI fragment from λ DA28-26, see Figure 1); lane L1, leaf mRNA hybridized to labeled λ FA28-5 (Fischer and Goldberg, 1982); lane L2, leaf mRNA hybridized to labeled pG26R5.4.

aries of the two domains, homologous regions greater than 2 kb in length are visible (see Figure 3, B and F). This result indicates that the two domains are related by partial homology over their entire length; that is, the two domains are homoeologous.

Figure 3, C and D, shows that heteroduplexes between DNA representing the linked *Gy*₁ and *Gy*₂ genes and that containing *Gy*₃ have a single-stranded 4.2-kb deletion/insertion loop. Comparison of these heteroduplexes with the restriction map shown in Figure 1 indicated that this loop represented the *Gy*₁ glycinin gene (2.65 kb) plus flanking DNA sequences (1.65 kb). Other heteroduplexed molecules were formed in which the *Gy*₁ and *Gy*₃ glycinin genes were hybridized, and in these cases a 4.3-kb single-stranded loop corresponding to *Gy*₂ was visualized (data

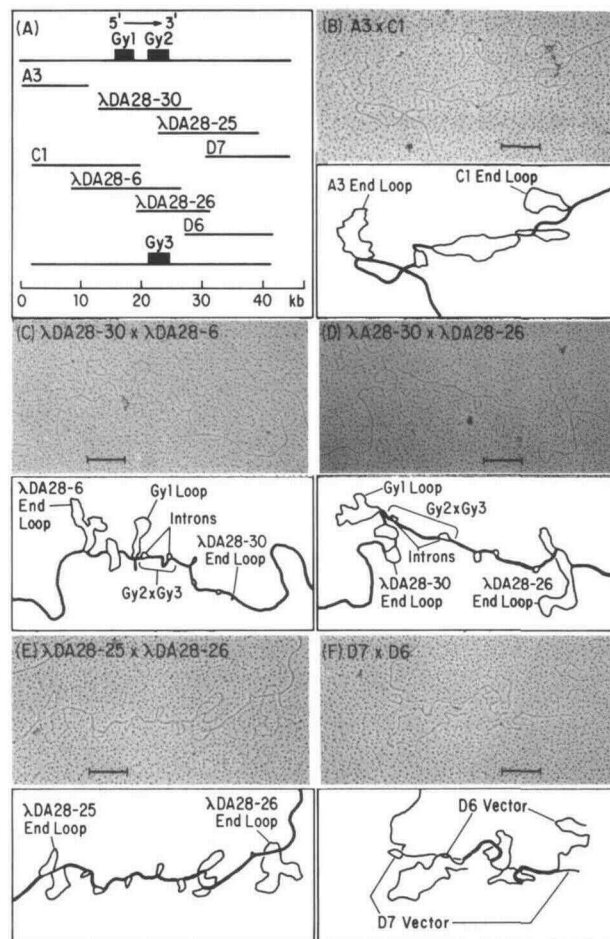


Figure 3. Heteroduplex Analysis of Cloned Glycinin Genes and Flanking Sequences.

Heteroduplexes were formed between phage DNAs, spread, and visualized by electron microscopy as described in "Methods". Bar equals 1 kb in all photographs. End-loops were formed by cloned sequences adjacent to vector sequences that were outside the overlap region.

(A) The organization and position of heteroduplexed phage DNAs relative to glycinin genes and each other.

(B) λ DA-3 $\times$ λ DC-1 heteroduplex. Approximately 45% of the overlap region was double-stranded.

(C) λ DA28-30 $\times$ λ DA28-6 heteroduplex. Approximately 48% of the overlap region was double-stranded.

(D) λ DA28-30 $\times$ λ DA28-26 heteroduplex. Approximately 47% of the overlap region was double-stranded.

(E) λ DA28-25 $\times$ λ DA28-26 heteroduplex. Approximately 52% of the overlap region was double-stranded.

(F) λ DD-7 $\times$ λ DD-6 heteroduplex. Approximately 30% of the overlap region was double-stranded. The orientation of the clones relative to the Charon 4 vector is such that when the soybean sequences form a heteroduplex, the vector sequences cannot hybridize and remain single-stranded.

not shown). Together these results indicate that the two domains are highly homologous, and that the *Gy₁/Gy₂* glycinin gene cluster differs from the *Gy₃* chromosomal domain by either a simple deletion or duplication of glycinin gene region of DNA.

Glycinin Genes Have Complex Structures

Glycinin gene structures were visualized by R-loop analysis using the electron microscope (Fischer and Goldberg, 1982). Figure 1 shows that the three Group-I glycinin genes contained four exons and three introns. The exon and intron sizes were calculated from the lengths of the paired and unpaired DNA strands (Figure 1, legend). By contrast, Figure 1 shows that the LF3 leaf gene R-loop is simple and does not contain intron structures visible by electron microscopy (Fischer and Goldberg, 1982).

To investigate glycinin gene structure in more detail, we sequenced the *Gy₁*, *Gy₂*, *Gy₃*, and *Gy₄* glycinin genes, as well as the 5'- and 3'-flanking regions of the *Gy₅* glycinin gene (Scallion et al., 1985; T.-J. Cho and N.C. Nielsen, unpublished data). Figure 4 summarizes these results, and Table 2 gives the percent nucleotide sequence homologies between the various regions in the genes. The actual sequences will be submitted to appropriate gene data bases or have been published previously (Scallion et al., 1985; Cho et al., 1989b). In agreement with the R-loop analysis, each of the five glycinin genes contains four exons and three introns. We established the positions of the introns in the sequences on the basis of (1) chemically determined amino acid sequence of the entire G2 subunit (Staswick et al., 1984), (2) comparisons with sequences of corresponding full length cDNA clones (Marco et al., 1984; Scallion et al., 1985; C.D. Dickinson, T.-J. Cho, and N.C. Nielsen, unpublished results), and (3) S1 nuclease protection experiments (Scallion et al., 1985; N.C. Nielsen, unpublished results). The three introns interrupted the coding regions at the same relative positions in each of five genes, but were variable in size. The sizes of the introns determined by nucleotide sequence analysis (Figure 4, legend) corresponded well with those determined by R-loop analysis (Figure 1, legend). Table 2 shows that, with the exception of the second intron in *Gy₁* and *Gy₂* (see "Discussion"), introns that corresponded to one another in each of the five genes exhibited little or no sequence homology (Table 2). As is typical of introns in other eukaryotic genes, the introns contained splice sites that obey the GT/AG cleavage rule (Breathnach et al., 1978).

Computer analysis of the gene sequences enabled the identification of several consensus sequences. First, TATA box sequences were located 25 bp to 30 bp upstream from transcription start sites. The latter were identified by S1 nuclease protection analysis with midmaturation stage embryo mRNA. Second, CAAT-like boxes were identified about 100 bp upstream from the transcription start site of

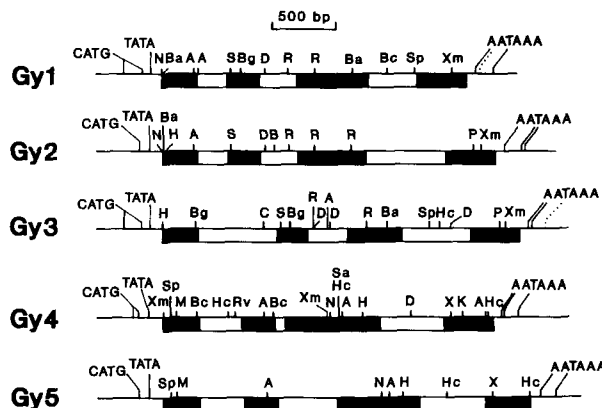


Figure 4. Detailed Restriction Maps for Five Glycinin Genes.

The nucleotide sequences either have been published (*Gy₄* [Scallion et al., 1985]), or will be published elsewhere (*Gy₁* [T.L. Sims and R.B. Goldberg, unpublished results]; *Gy₂* [V.H. Thanh, B.J. Scallion, and N.C. Nielsen, unpublished results]; *Gy₃* [T.-J. Cho and N.C. Nielsen, unpublished results]). The *Gy₅* intron, exon, and flanking region sizes were obtained either by incomplete sequence analysis of λ FG11 or S1 nuclease protection experiments (T.-J. Cho and N.C. Nielsen, unpublished results). All nucleotide sequences are available upon request. Intron and exon sizes are (Exon1:Intron1:Exon2:Intron2:Exon3:Intron3:Exon4): *Gy₁*, 286:228:254:291:558:381:390; *Gy₂*, 277:238:254:292:537:624:390; *Gy₃*, 286:617:245:312:525:439:390; *Gy₄*, 289:332:266:75:747:501:390; *Gy₅*, 292:350:263:470:645:520:351. Abbreviations: A, AclI; B, BamHI; Ba, BstI; Bc, BclI; Bg, BglII; C, ClaI; D, DraI; H, HindIII; Hc, HincII; K, KpnI; M, MluI; N, NcoI; P, PvuI; R, EcoRI; Rv, EcoRV; S, SmaI; Sa, Sall; Sp, SphI; X, XhoI; Xm, XmnI.

each gene, although correspondence between these sequences and the animal consensus sequence (Benoist et al., 1980) was poor. Only 4 bp out of 9 bp were homologous with the animal consensus sequence in the two Group-II glycinin genes (*Gy₄* and *Gy₅*), whereas the corresponding positions in the Group-I glycinin genes (*Gy₁*, *Gy₂*, *Gy₃*) had even less identity. Third, seed protein-specific consensus sequences were identified in the 5' regions of the glycinin genes. These included the 8-bp 5'-CATGCATG-3' RY sequence (Dickinson et al., 1988), and the 9-bp 5'-CAACACAAT-3' sequence (Goldberg, 1986). Finally, multiple 5'-AATAAA-3' polyadenylation sequences (Fitzgerald and Shenk, 1981) were located in the 3'-flanking regions of the five glycinin genes (Figure 4), and 5'-AACAAUGGC-3' consensus sequences were located at the gene translation initiation sites (Lutcke et al., 1987). The latter were not as well conserved in the Group-II glycinin genes as they were in the Group-I genes. Together these results indicate that the glycinin gene structures are highly conserved, and that potential developmental control sequences are present in the 5' regions of the genes.

Table 2. Percent Similarities in Various Regions of Glycinin Genes

Region	Gy ₁ /Gy ₂	Gy ₂ /Gy ₃	Gy ₁ /Gy ₃	Gy ₁ /Gy ₄	Gy ₃ /Gy ₄	Gy ₂ /Gy ₄	Gy ₄ /Gy ₅
5'-Flanking ^a	84	79	90	61	58	66	91
Leader	80	68	72	42	42	45	81
Exon 1	93	92	95	61	62	56	94
Intron 1	49	50	47	46	45	44	— ^b
Exon 2	91	90	93	66	65	63	92
Intron 2	99	80	83	44	45	44	—
Exon 3	86	87	91	63	65	54	90
Intron 3	56	54	46	44	41	45	—
Exon 4	93	94	92	51	51	50	85
3'-Flanking	65	89	68	40	32	35	79

^a The 5'-flanking region refers to 200 bp upstream from the transcription start site. The untranslated leader refers to the region from transcription start to translation start. The 3'-flanking region refers to 200 bp following the translation stop codons.

^b Sequence of Gy₅ introns not determined.

Glycinin Subunit Sequences Are Derived from the Glycinin Genes

The amino acid sequences deduced from each of the five glycinin genes are presented in Figure 5, and Table 3 gives the percent sequence identities between the translated glycinin proteins. Also shown for comparison are percent sequence identities between the glycinin subunits and a number of related 11S proteins that have been reported. With the exception of G5, the glycinin primary structures were derived from the nucleotide sequence of genomic clones from the variety Dare. The data for G5 originated from previously published cDNA sequences (Fukazawa et al., 1985; Scallan et al., 1985). The deduced primary structures are in good agreement with ones derived from NH₂-terminal sequence analysis (Moreira et al., 1979, 1981) and permit unambiguous identification of the corresponding genes. In the case of G2, the entire subunit has been sequenced at the amino acid level (Staswick et al., 1984) and differs from the DNA derived sequence by only 2 out of 558 residues. As can be seen from the comparisons given in Table 3, the glycinin subunits can be separated into two distinct subfamilies based on percent sequence homology. As predicted earlier from amino acid sequence determinations (Nielsen, 1984), one family includes the G1, G2, and G3 glycinin subunits, and the other contains G4 and G5. Homologies between members of the same subfamily ranged from 80% to 90%, but percent identities between members of different groups were less than 50% (Table 3). Similar results were obtained when the nucleotide sequences of corresponding exons of the five genes were compared (Table 2).

Figure 5 shows that the carboxyl ends of the acidic chains of glycinin subunits were especially divergent. Because of this property, we refer to these ends as the hypervariable region, or HVR. One of the striking features of the HVR is the high content of negatively charged

aspartate and glutamate amino acids. The aspartate/glutamate-rich repeats could have evolved by repeated duplication of a glutamate or aspartate codon. As shown in Figure 6, a larger region containing the reiterated acidic sequence has been duplicated, once in the case of the G5 glycinin subunit and twice in the case of the G4 subunit (compare Figures 5 and 6). These duplications account for the size differences between glycinin subunits (Table 1).

Glycinin Genes Are Coordinately Expressed

We investigated whether individual glycinin gene family members were coordinately expressed during seed development by hybridizing cDNA clones representing each of the glycinin family members with dot-blots containing mRNA isolated from different embryonic stages, from postgermination cotyledons, and from the mature plant leaf, root, and stem. We also used the pG3-1 cDNA clone that represented non-Group-I and non-Group-II glycinin-related genes. To minimize cross-hybridization between family members, stringent hybridization conditions and gene-specific probes were employed.

Figure 7A shows that all of the glycinin mRNAs accumulated and decayed in coordinated patterns during embryogenesis. Glycinin mRNAs were first detected 35 days after flowering, reached maximal levels 70 days to 87 days after flowering, and then decayed to nondetectable levels by 100 days after flowering. Glycinin mRNAs were not detected in postgermination cotyledons or mature plant leaves, roots, and stems (Figure 7A). Figure 7A also shows that the pG3-1 cDNA clone reacted with mRNAs that were less abundant in embryos than those encoding the Group-I and Group-II glycinin polypeptides, but that they accumulated and decayed during embryogenesis with kinetics that were similar to those for glycinin mRNAs. Figure 7B shows that, at the level of DNA/DNA hybridization, our probes detected gene-specific sequences. Together, these

	signal									
G1	MAK....LV	FSLCFLFSG	CCFAFSREQ	<u>PQNECQIQK</u>	<u>LNALKPDNRI</u>	ESEGLIETW	NPNNKPFQCA	GVALSRCTLN	RNALRRPSYT	90
G2	MAK....LV	LSLCFLFSG	C.FAL..REQ	<u>AQNECQIQK</u>	<u>LNALKPDNRI</u>	ESEGGFIETW	NPNNKPFQCA	GVALSRCTLN	RNALRRPSYT	
G3	MAK....LV	LSLCFLFSG	CCFAFSREQ	<u>PQNECQIQK</u>	<u>LNALKPDNRI</u>	ESEGGFIETW	NPNNKPFQCA	GVALSRCTLN	RNALRRPSYT	
G4	MGKPF.TLSL	SSLCLLLSS	ACFAISS...	<u>SKLNECQLNN</u>	<u>LNALPDHRV</u>	ESEGLIQTW	NSQHPELKCA	GVTVSKLTN	RNGLHLPSYS	
G5	MGKPF.TLSL	SSLCLLLSS	ACFAITS...	<u>SKFNECQLNN</u>	<u>LNALPDHRV</u>	ESEGLIETW	NSQHPELQCA	GVTVSKRTLN	RNGSHLPSYL	
	acidic									
G1	NGPQEIIYQ	GKGIFGMIYP	*GCPSTFEEPQ	Q....PQQRG	QSSRPQDRHQ	KIYNFREGDL	IAVPTGVAWW	MYNNEDTPVV	AVSIIDTNSL	180
G2	NGPQEIIYQ	GNGIFGMIYP	GCPSTYQEPQ	E....SQQRG	RSQRPQDRHQ	KVHRFREGDL	IAVPTGVAWW	MYNNEDTPVV	AVSIIDTNSL	
G3	NAPQEIIYQ	GSIGFMIYP	GCPSTFEE..	PQQRG	QSSRPQDRHQ	KIYHFREGDL	IAVPTGFAYW	MYNNEDTPVV	AVSLIDTNSF	
G4	PYPRIIIAQ	GKGLGVAIP	GCPTFEEPQ	EQSNRRGSR	<u>QKQLQDSHQ</u>	<u>KIRHFNEGDV</u>	LVIPPGVPYW	TYNTGDEPVV	AISLDTSNF	
G5	PYPQMIIVVQ	GKGAIGFAFP	GCPTFEKPQ	QSSRRGSR	Q.QQLQDSHQ	KIRHFNEGDV	LVIPLGVPYW	TYNTGDEPVV	AISPLDTSNF	
	HVR									
G1	VKGGLSVIKP	PTDEQQRRPQ	EEEEEEDEK							360
G2	VKGGLRVTP	AMRKPPQ...	EEDDDDEEQ							
G3	VKGGLSVISP	PTEEQQRPE	EEE.....K							
G4	VEGGLSVISP	KWQEQQDEDE	DEDEDEDEQ	IPSHPPRRPS	HGKREQDEDE	DEDEKPRPS	RPSQGKREQD	QDQDEDEDE	EDQPRKSREW	
G5	VEGGLSVISP	KWQEQQDEDE	DEDEE..YGR	TPSYPPRRPS	HGKHEDDEDE	DEEDQPRPD	HPPQ.....			
	basic									
G1	PQCKGDKHC	QPRGGSQSK	SRRNGIDETI	<u>CTMRLRHNI</u>	QTSSPDIYNP	QAGSVTTATS	LDFPALSRL	LSAEFGSLRK	NAMFVPHYNL	
G2	PQCVETDKGC	QR....QSKR	S.RNGIDETI	<u>CTMRLRHNI</u>	QNSSPDIYNP	QAGSITTATS	LDFPALWLLK	LSAQYGLSRK	NAMFVPHYTL	
G3	PDCDEKDKHC	QSQ.	S.RNGIDETI	<u>CTMRLRHNI</u>	QTSSPDIFNP	QAGSITTATS	LDFPALSRL	LSAQFGSLRK	NAMFVPHYNL	
G4	RSKKTQPRRP	RQEPRERGC	ETRNGVEENI	<u>CTMLKHENIA</u>	RPSRADFYNP	KAGRISTLNS	LTLPALRQFQ	LSAQYVVLKY	NGIYSPHWNL	
G5	RPSRP	EQQEPRGRGC	QTRNGVEENI	<u>CTMLKHENIA</u>	RPSRADFYNP	KAGRISTLNS	LTLPALRQFG	LSAQYVVLKY	NGIYSPDWNL	
	480									
G1	NANSIIYALN	GRALIVQVNC	NGERVFDGEL	QEGRVLIYPQ	NFVVAARSQS	DNFEYVSFKT	NDTPMIGTLA	GANSLLNALP	EEVIQHTFNL	
G2	NANSIIYALN	GRALVQVNC	NGERVFDGEL	QEGGVLIYPQ	NFAVAARSQS	DNFEYVSFKT	NDRPSIGNLA	GANSLLNALP	EEVIQHTFNL	
G3	NANSIIYALN	GRALVQVNC	NGERVFDGEL	QEGQVLIYPQ	NFAVAARSQS	DNFEYVSFKT	NDRPSIGNLA	GANSLLNALP	EEVIQHTFNL	
G4	NANSVIYVTR	GQGVVQVNC	QGNVAVFDGEL	RRGQLLVVPQ	NFVVAEQAGE	QGFYIVFKT	HHNAVTSYL.	..KDVFRVIP	SEVLASYNL	
G5	NANS.VTMTR	GKGRVQVNC	QGNVAVFDGEL	RRGQLLVVPQ	NPAVAEQGGE	QGLEVVVFKT	HHNAVSSYI.	..KDVFRVIP	SEVLASYNL	
	573									
G1	KSQARQIKN	NNPFKFLVPP	QESQ....KR	AVA	573					
G2	KSQARQVKN	NNPFSFLVPP	QESQ....RR	AVA						
G3	RQSQVSELKY	EGNWGLVNP	.ESQGSFPRV	KVA						
G4	RRQARQVKN	NNPFSFLVPP	KESQ....RR	VVA						
G5	GQSQRQLKY	QNSGSLVNP								

Figure 5. Alignment of Derived Glycinin Amino Acid Sequences.

Gaps have been introduced to maximize alignments (Needleman and Wunsch, 1970). Arrowheads indicate major sites of posttranslational modification. Cysteine residues marked by an asterisk (*) are involved in interchain linkages, whereas those marked by a cap (^) identify conserved residues possibly involved in intrachain linkages. Doubly underlined residues indicate amino acids that follow cleavages shown to occur in soybean cultivar CX635-1-1 (Staswick et al., 1984). Residues that agree with amino acid sequence data are underlined, except in the case of G2 where the disagreements are underlined. The position of the glycinin hypervariable region (HVR) is indicated.

results indicate that glycinin gene expression is highly coordinated during development, and that glycinin-related genes encode low prevalence glycinin mRNAs.

We hybridized ^{32}P -labeled nuclear RNAs synthesized *in vitro* to determine whether the coordinated accumulation of glycinin mRNAs was due to transcriptional or posttranscriptional regulatory processes. Figure 8 shows that the transcription of all three glycinin genes was first detectable 25 days after flowering, reached a maximal level 55 days after flowering, and then declined to a low level by 91 days after flowering. A comparison of the results presented in Figures 7 and 8 shows that increases and decreases in glycinin mRNA prevalences paralleled similar changes in the relative transcription rates of individual genes. Significantly, Figure 8 also shows that there was not an absolute correlation between relative gene transcription rates and mRNA prevalences. For example, 35 days after flowering, the relative transcription rate of the pG3-1 glycinin-related sequence was higher than that of the G2 and G5 glycinin genes, although the pG3-1 G* mRNA prevalence was lower than that of the Gy_2 and Gy_5 mRNAs. Together these results indicate that glycinin gene transcription and mRNA accumulation are highly coordinated during embryogenesis, and that both transcriptional and posttranscriptional processes are important in regulating glycinin and glycinin-related mRNA levels.

DISCUSSION

Glycinin Gene Regions Are Unlinked and Contain Non-Seed Protein Genes

The overlapping genomic clones shown in Figure 1 define two chromosomal domains that span a total of about 85 kb of the soybean genome. Two glycinin genes, Gy_1 and Gy_2 , are linked in the same orientation about 1.5 kb apart in one domain, whereas the Gy_3 glycinin gene is located in the second domain. The Gy_1/Gy_2 cluster and the Gy_3 gene are both linked to restriction fragments that hybridize weakly to glycinin cDNA clones pG23 and pG258 (Scallan et al., 1985). Because we have not sequenced these fragments, we do not know how they relate to glycinin genes. However, because these fragments hybridize specifically with clones pG23 and pG258, they are more homologous with Group-II glycinin genes than those in Group I. In addition, several other restriction fragments in the two domains hybridized specifically with leaf mRNAs (Figure 2A), and did not react detectably with embryo mRNA (Figure 2B). They presumably contain genes expressed in leaves but not embryos. The close linkage between seed protein genes and non-seed protein genes with different expression programs also occurs in the soybean lectin gene region (Okamuro et al., 1986) and appears to be a general feature of the soybean genome.

Table 3. Percent Sequence Identity between 11S Subunits

Subunit	G1	G2	G3	G4	G5	LA ^a	LB ^b	C1 ^c	C2 ^d	Cr ^e	Gl ^f	Oa ^g
G1		84	89	48	46	64	47	42	38	41	37	34
G2	84		86	46	45	66	44	43	39	38	38	34
G3	89	86		49	47	67	45	43	39	41	38	35
G4	48	46	49		87	44	63	37	37	37	39	37
G5	46	45	47	87		42	62	37	37	38	38	36
LA ^a	64	66	67	44	42		40	37	40	38	37	35
LB ^b	47	44	45	63	62	40		43	43	39	41	37
C1 ^c	42	43	43	37	37	37	43		46	44	43	38
C2 ^d	38	39	39	37	37	40	43	46		43	45	40
Cr ^e	41	38	41	37	38	38	39	44	43		40	40
Gl ^f	37	38	38	39	38	37	41	43	45	40		66
Oa ^g	34	34	35	37	36	35	37	38	40	40	66	

^a *LegA* gene from *Pisum sativum* (Lycette et al., 1984).

^b *LegB4* gene from *Vicia faba* (Baumlein et al., 1986).

^c cDNA C-94 from *Gossypium hirsutum* (Chlan et al., 1986).

^d cDNA C-134 from *G. hirsutum* (Chlan et al., 1986).

^e Cruciferin cDNA from *Brassica napus* (Simon et al., 1985).

^f Gutellin cDNA from *Oriza sativa* (Takaiwa et al., 1986).

^g Globulin from *Avena sativa* (Shotwell et al., 1988).

Like the glycinin genes and the G* restriction fragments, the leaf genes occupy analogous positions in the two domains (Figures 1 and 2). The heteroduplexes shown in Figure 3 indicate that regions of DNA sequence homology extend over the entire 42 kb in each of the two domains, and that the Gy_1/Gy_2 cluster differs from the Gy_3 gene domain by a simple 4.3-kb deletion or duplication corresponding to the Gy_1 gene.

In a previous study we showed that the Gy_1 , Gy_2 , and Gy_3 glycinin genes in the two domains were not alleles of the same genetic locus (Fischer and Goldberg, 1982). The genetic analysis described in the accompanying paper confirms this conclusion and establishes that the Gy_1/Gy_2 and Gy_3 regions segregate independently from each other and from regions containing the Group-II glycinin genes (Cho et al., 1989a). It is possible that the two Group-I glycinin chromosomal domains evolved as a consequence of a large chromosomal duplication. If the two are on the same chromosome, however, the duplicated domains must be separated by at least 50 map units for them to segregate independently. The two domains could also be the result of a global duplication, and be located on homoeologous chromosomes of an ancient tetraploid (Palmer, 1978; Cho et al., 1989a). Lee and Verma (1984) proposed tetraploidization as a mechanism to account for the evolution of the two complex genetic loci in soybeans, and Okamura et al. (1986) showed that regions flanking the soybean lectin gene were present in one or more other locations in the soybean genome. Our data are therefore consistent with previous observations that the soybean genome contains homoeologous chromosomal domains.

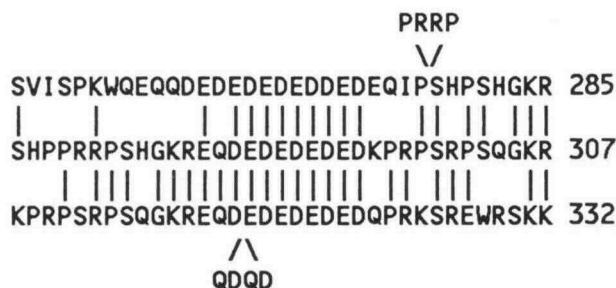


Figure 6. Evolution of the Glycinin Hypervariable Region.

Internal repeats in the G4 hypervariable region have been aligned to show a large coding segment of the gene was duplicated. Bars indicate sequence similarities. Numbering is from the NH₂ terminus of the acidic polypeptide (Figure 5).

The Glycinin Gene Family Is Organized into at Least Two Groups

We sequenced the five genes denoted *Gy*₁ to *Gy*₅ (Figure 4). Each gene encodes one of the prevalent subunits that constitute a part of glycinin hexamers stored in dormant seeds (Nielsen, 1985). The primary structures of subunit precursors deduced from the genes match the amino acid sequences of purified acidic and basic glycinin polypeptides (Staswick et al., 1984). Two other glycinin-related sequences, denoted G* (Figures 1, 7, and 8), were identified that were complementary to 2.1-kb seed-specific mRNAs. These included sequences homologous with the pG3-1 cDNA clone and those related to the G* DNA fragments in glycinin DNA phages (Figure 1). Sequences represented by pG3-1 appear to be related to Group-I glycinin genes, whereas those represented by G* phage DNA fragments are more homologous with the Group-II genes. These results suggest that other glycinin genes are present in the soybean genome that have diverged significantly from the Group-I and Group-II gene subfamilies.

Figure 7 shows that the pG3-1 glycinin-related gene sequences hybridized with mRNAs that accumulate to a low abundance in the seed. Proteins encoded by these mRNAs are expected to be present in minor amounts in the seed compared to subunits produced at the direction of the five prevalent glycinin genes. Lei et al. (1983) identified a number of polypeptides in soybeans that appeared to be related to the major glycinin peptides, and were present in low amounts in the seed. Although these minor polypeptides could represent modified versions of the prevalent glycinin subunits, they could also reflect proteins encoded by glycinin-related genes. A complex mixture of major and minor 11S subunits would be anticipated if the glycinin-related genes encoded other subunit families. If so, this situation would be similar to that reported in pea, where a number of major and minor acidic and basic

polypeptide chains of legumin have been identified by two-dimensional electrophoresis (Matta et al., 1981; Gatehouse et al., 1986).

Glycinin Genes Have Similar Structures

Figure 4 shows that the five glycinin genes share a common structure and have coding regions that are interrupted by introns three times. The presence of three introns is conserved in the homologous legumin-like genes of pea (Lycett et al., 1984), rice (Takaiwa et al., 1986, 1987), and oat (Shotwell et al., 1988). Homologous genes in the bean *Vicia faba* (LeB4, Baumlein et al., 1986) and sunflower (T. Thomas, personal communication) are the only 11S genes that have been reported so far that do not contain three introns. In these genes the second and third introns are analogous to those present in glycinin genes, but the first intron is missing.

The nucleotide sequence of the introns in glycinin genes is not highly conserved, especially between members of different glycinin subunit groups (Table 2). One notable exception is the 99% sequence identity between the second introns of the *Gy*₁ and *Gy*₂ glycinin genes. Because intron-2 in *Gy*₁ and *Gy*₂ is less similar to that of *Gy*₃, and is

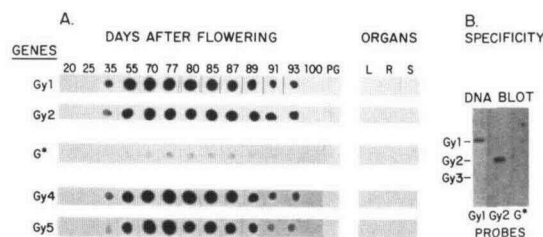


Figure 7. Changes in mRNA Concentration from Glycinin and Glycinin-Related Genes during Development.

(A) Dot-blot analysis. mRNAs from developing embryos (0.2 µg), postgermination cotyledons (0.1 µg), leaves (0.6 µg), roots (1.0 µg), and stems (0.2 µg) were immobilized on nitrocellulose filters and hybridized with sequence excess amounts of ³²P-labeled DNA. Probes used were: *Gy*₁, pG1-2 (0.7-kb insert); *Gy*₂, pG2-1 (0.7-kb insert); G*, pG3-1 (0.3-kb insert); *Gy*₄, pG258 (0.7-kb insert); *Gy*₅, pG23 (0.6-kb insert). The origin of these clones is given in "Methods". The specific activity of all labeled DNAs was 3 × 10⁸ cpm/µg. Wash conditions were 1 M Na⁺, 50% formamide, 60°C, which is about 10°C below the melting temperature of perfectly matched hybrids (Bletz et al., 1983). The autoradiograms were exposed for 2 hr.

(B) Specificity of probes. To demonstrate that the *Gy*₁, *Gy*₂, and G* probes were specific for their respective family members, 400 pg of EcoRI-digested λDA28-30 (*Gy*₁/*Gy*₂, Figure 1) and 400 pg of λDA28-26 (*Gy*₃, Figure 1) were separated by electrophoresis in the same lane and blotted to nitrocellulose. They were hybridized to ³²P-labeled DNA as in (A). Probes used were: *Gy*₁, pG1-2; *Gy*₂, pG2-1; G*, pG3-1; *Gy*₅, pG23.

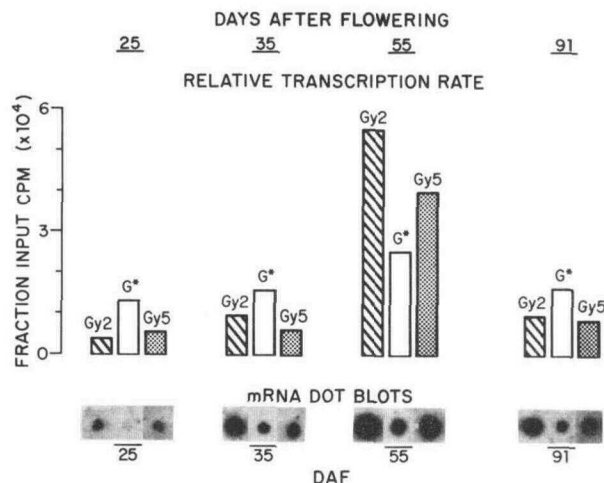


Figure 8. Transcription of Glycinin and Glycinin-Related Genes.

Excess cloned soybean DNA sequences (1.6 μ g) were immobilized onto nitrocellulose discs and hybridized to ^{32} P-labeled nuclear runoff RNA synthesized in vitro using embryo nuclei isolated 25, 35, 55, and 91 days after flowering (see "Methods"). The immobilized DNAs were: *Gy*₂, pG2-1; *G**, pG3-1; *Gy*₅, pG23. Input counts per minute were: 25-day nuclei, 1.77×10^7 cpm; 35-day nuclei, 1.84×10^7 cpm; 55-day nuclei, 1.08×10^7 cpm; 91-day nuclei, 1.16×10^7 cpm.

only 44% identical to intron 2 of *Gy*₄, it is unlikely that the homology has functional significance. More likely, this close sequence identity is a vestige of recent gene conversion between the two linked genes. Such events are common among multigene families (Slightom et al., 1980) and complicate deduction of evolutionary relationships.

Glycinin Subunits Have a Common Protein Structure

Comparison of the five subunits shown in Figure 5, and summarized in Table 3, supports our earlier suggestion that there are at least two glycinin protein subfamilies (Nielsen, 1984). Two subfamilies of legumin-like proteins are also present in pea (Domoney and Casey, 1985) and the bean *Vicia faba* (Baumlein et al., 1986). Table 3 shows that the LegA subunit of pea is more similar to Group-I glycinin subunits than those in Group II. The opposite is true for the LeB4 subunit of the bean *V. faba*. It is interesting that the level of identity between Group-I glycinin subunits and the legA subunit of pea is higher than the between group similarities among glycinin subunits. This suggests that the divergence which gave rise to the two groups occurred prior to speciation of soybean, pea, and the bean *V. faba*. The legumin-like cotton β -globulins are also divided into two families (Chlan et al., 1986). However,

the divergence between the glycinin proteins and those of cotton is too great to speculate about evolutionary relationships among these various legume-like proteins.

The presence of subfamilies of 11S proteins raises questions regarding the physiological significance, if any, of the subunit heterogeneity that is observed. Are subunits from different subfamilies functionally interchangeable in seed oligomers, or do certain subunits in each of the two families perform different essential functions during assembly of glycinin hexamers? An answer to this question could have important implications for the genetic engineering of these proteins. Two cultivars of *Glycine max* have been identified that lack a different glycinin subunit (*G*₄, Staswick and Nielsen, 1983; *G*₃, Cho et al., 1989b). Although individual subunits are missing in these two cultivars, 11S complexes are still produced and their seeds contain normal levels of glycinin protein (Kitamura et al., 1984). This result suggests that there is genetic polymorphism with respect to glycinin subunit composition, and that subunit heterogeneity may not have a functional relevance. Recently, evidence has been obtained for functional differences between the two groups of glycinin subunits with respect to their assembly properties in vitro (Nielsen, 1989). Together, these results indicate that, although there is not a strict stoichiometric relationship between glycinin subunits for oligomer assembly, structural differences do exist between subunits that affect assembly of glycinin hexamers.

Figure 5 shows aligned amino acid sequences of five glycinin subunits, and Figure 9 compares the *G*₄ glycinin sequence with those of a number of other legumin-like proteins for which complete sequences are available. Conserved asparagine-glycine amino acids are located at the site in the precursor proteins where cleavage occurs to form the acidic and basic chains that constitute mature 11S subunits. To date, the protease that catalyzes the cleavage at this site has not been described. It is worth noting, however, that asparagine-glycinin peptides are unique in that they are capable of forming a cyclic imide (Bornstein and Balian, 1977). It is possible that such a cyclic structure could be involved in the proteolytic cleavage mechanism. Although other asparagine-glycine peptides exist in the glycinin subunits that are not cleaved, these amino acids are probably buried in the protein and are not accessible to the protease.

A single cystine covalently links the acidic and basic polypeptide chains in *G*₁, *G*₂, *G*₄, and *G*₅ (Staswick et al., 1984). This is also likely to be the case for *G*₃ because these cysteine residues are strictly conserved (Figure 5, positions 112 and 390). Figure 9 shows that the interchain cystines are also conserved in other legumin-like subunits for which amino acid sequences are available (Figure 9, positions 128 and 448). Only two other cysteine residues in the glycinin subunits are as strictly conserved in the subunits (Figure 5, positions 36 and 69), and Figure 9

G4	MGKPFTLSLS	.SLCLLLSS	ACFAISSSKL			N	ECQLNNLNL	EPDHRVESEG	GLIQTWNSQH	PELKACGVTV	90
LA	MAKLLALSLS	..FCFLLGG	.CFALREQPQ	Q.....		N	ECQLERLDAL	EPDNRIESEG	GLIETWNPNN	KQFRCAGVAL	
C1	VNPSL	LFLSLLFLFN	GCLARQTFFS	QQSQ.....		N	ECQINRLRAS	APQTRIRSEA	GTTEWNPNC	QQLRCAGVSV	
C2	MAYTSL.SF	SV.CLLVLFH	GCCAQIDLVT	NHHQDPPWGO	PQQPQPRHQS	QCQLQNLNL	QPKHRFRSEA	GETEFWDQNE	DQFQCAGVAF		
Cr	MARLSSLSF	SLALLIFL.H	GSTAQQFP..			N	ECQLDQLNAL	EPSHVLKAE	GRIEVDHHA	PQLRCAGVSV	
GI	MASINRPVIF	FTVCLFLLCN	GSQAQQLLGQ	STSQAQSSRR	GSP.....	R	ECRFDRQLAF	EPIRSVRSQA	GTTEFFDVSN	EQFQCTGVSV	
Cons	MA----	L-SF	--LCLLLL--	GC-AQQ-L--	---Q-----	N	ECQLDRLNAL	EP-HR--SEA	G--EFW--N-	-Q-RCAGVS-	
VR1											
G4	SKLTNRNGL	HLPSYSPYPR	MIIIAQCKGA	LGVAIPEGCE	TFEEPQESN	RR.....				GSRSQK	180
LA	SRATLQRNAL	RRPYSSNAPQ	EIFIQQNGY	FGMVFPGCPE	TFEEPQESE.					QGEG	
C1	MRQTIEPNGL	VLPSTNAPQ	LLYIVQGRGI	QGIVMPGCAE	TFQDSQQWQH	QSRGR.....				IQD	
C2	LRHKIQKGL	LLPSFTSAPM	LFYVEQGEI	HGAVFPGCPE	TYQSQQQN.					IQD	
Cr	VRYIESKGL	YLPSTSTAK	LSFAKGEGL	MGRVVL.CAE	TFQDSVVFQ	SGGSPFEGGQ	GGGQQGQGGH	GGGQQGQGGQ	GGGQQGQGGQ	GGGQQGQGGQ	
GI	VRRVIEPRGL	LLPHYTNAS	LVYIIQGRGI	TGPTFPGCPE	SYQQQFQSG	QAQLTES...				QSQS	
Cons	-R-TIERNGL	-LPS-TNAP-	LIYI-QG-GI	-G-VFPGCPE	TF-E-QQQ--	-----	-----	-----	-----	-QSQ-	
VR2											
G4	QQLQDSHQKI	RHFNEGDVLV	IPGPVYWTY	NTGDEPVVAI	SLLDTSNFIN	QLDQTPRVFY	LAGNPDIEYP	ETMQQQQQK	SH.....	270	
LA	RRYRDRHQKV	NRREGDIIA	VPTGIVFWY	NQDQTPVIAV	SLLDIRSSNN	QLDQMPRRFY	LAGNHEQEF.				
C1	..FQDQHQKV	RRFRQGDIIA	LPQGVVHWSY	NDGNERVVTI	NLLDTGNSAN	QLDNIPRRFH	LAGNPEEQR	QLRLAQMQ	GRSERGESE		
C2	RPQRDRHQKL	RLKEGGVVA	LPAGVAHWIF	NNGRSQDLVL	ALVDVQNDAN	QLDENFRKFF	LAGSPQGGV	RGGQSRDRNQ	RQSRTRQGER		
Cr	QGFDRMHQKV	EHIRTGDTIA	THPGVAQWY	NDGNQPLVIV	SVLDLASHQN	QLDRNPRPFY	LAGNPPQG..				
GI	QKFDEHQKI	HRFRQGDVIA	LPAGVAHWY	NDGEVPPVAI	YVTDLNNGAN	QLDPRQRDFL	LAGNK.....				
Cons	Q-FRD-HQKV	RRFRQGDVIA	LP-GVAHW-Y	NDGD-PPVA-	SLLD--N-AN	QLD--PR-FY	LAGNPE-E--	-----	-----		
G4	.GGKQGGHQ	QEEEEEGGSV	LSGFSKHFLA	QSFNTNEDIA	EKLQSP..DD	ERKQIVTEG	GLSVISPKWQ	EQQDEDEDED	EDDEDEQIPS	360	
LA	LQYQHQGGK	QEQENEGNNI	FSGFKRDFLE	DAFNVRNHRIV	DRLQGRNEDE	EKGAIKVKKG	GLSIIISPPEK	QARHQGRSQ	EEDDEEEKQP		
C1	EEEGEGEE	EEDNPSRRSR	HQEEEEGGRE	SSSCNLLCA	FDRNFLAQAF	NVDHDIIRKI	QRVRGNRGTI	IRVRDRLQV	TPRMEEER		
C2	E.....EE	ESQESGGNNV	LSGFRDNLIA	QAFGIDTRLA	RKLQNE..RD	NRGAIVRMEH	GFEPPEEGQR	RQGREEEGEE	EREKQWRRQ		
Cr	QVWIE	GREQQPKNI	LNGFTPEVLA	KAFKIDVRTA	QQLQNG..QD	NRGNIIRVQG	PFSVIRPPLR	SQRPEE...			
GI	...RNPQAYR	REVEERSONI	FSGFSTELLS	EALGVSGQVA	RQLQCC..ND	QRGEIVRVEH	GLSLQPYAS	LQEQEQGGVQ	SRERYQEGQY		
Cons	-----Q---E	-EEE---NI	-SGF--E-LA	-AF--N--IA	--LQ-----D	NRG-IVRVEG	GLSVI-P---	-Q--EEE---	E---EE---		
HVR											
G4	HPPRRPSHGK	REQDEDED	EDKPRPSRPS	QKREQDQDQ	DEDEDDEDQ	PKRSREWRK	KTQPRRPRQE	EPRERGCTER	NGVEENICTL	450	
LA	RHQRSRQEE	EEDDFERQP	RHQRRRGEE	EEDKKERGS	QKGSRRQG.				NGLEETVCTA		
C1	EERQEQRYR	HTRGGSQ...							NGLEETFCSM		
C2	ESQEEGSEEE	EREERGRRR	RS.....						NGLEETFCSM		
Cr									NGLEETICSA		
GI	QSQYGSQC.								NGLEETFCCT		
Cons	-----S---	---E-----	-----	-----	-----	-----	-----	-----	NGLEET-C--		
G4	KLHENIARPS	RADFYNPAG	RISTLNSLT	PALRQFQLSA	QYVVLKNGI	YSPHWNLNAN	SVIYVTRGQG	KVRVNVCCQN	AVFDGELRRG	540	
LA	KLRNLIGPSS	SPDIYNPEAG	RIKTVTSLD	PVLRLWLKLSA	EHGSLHKNAM	FVPHYNLNAN	SIYALKGRA	RLQVNVCCQN	TVFAGELEAG		
C1	RIKENLADPE	RADIFNPAG	RISTLNRFLN	PILQRLLSA	ERGVLYNRAG	LIPQWNVNAH	KILYMLRGCA	RVQVNVHNGD	AVFDNDVQGG		
C2	RLKHR.TPAS	SADVFNPARG	RITTVNSFNL	PILQYLQLSA	ERGVLYNNAI	YAPHWNMNAH	SIVYITRGNG	RIQIVSENGE	AIFDEQVERG		
Cr	RCTDNLDDPS	NADVYKPLG	YISTLNSYDL	PILRFLRLSA	LRGSIRQNAM	VLPQWNNANAN	AVLYVTDGEA	HVQVNVNDNG	RVFDGQVSQ		
GI	RVRQIDNP	RADTYNPRAG	RVTNLNTQNF	PILSLIQMSA	VKNLYQNAL	LSPFWNINAH	SVYITQGRA	RVQVNVNNGK	TVFNGELRRG		
Cons	RL-ENI--PS	RAD-YNP-AG	RISTLNS-NL	PILR-LQLSA	ERGVLY-NAM	--PHWNLNA-	S--Y-TRG-A	RVQVNV-NGD	AVFDGE--RG		
G4	QLLVPPQNFV	VAEQAG.EQG	FEYIVFKTH	NAVTSYKLD.	..VFRAIPSE	VLAHSYNLRQ	SQVSELKYEG	NWGPLVPES	QQGSPRVKVA	630	
LA	RALTVPQNYA	VAAKSLSDR.	FSYVAFKTD	RAGIARLAGT	SSVINNLPLD	VVAATFNLQR	NEARQLKSNN	PFKFLVPARE	SENRAA...		
C1	QLLVPPQNFA	FMKQAGNE.G	AEWISFTNS	EATNTPMAGS	VSMFRALPEE	VVAASYQVSR	EDARRIKFNN	KNTFFFTPSQ	SETTADA...		
C2	QVITVPQNH	VVKKAG.RRG	FEWIAFKTNA	NAKISQIAGR	VSIMRGLPVD	VLANSFGISR	EEAMRLKHNR	QEVSVFSPRQ	GSQK.....		
Cr	QLLSIPQGF	VVKRATSEQ.	FRWIEFKTNA	NAQINTLAGR	TSVLRGLPLE	VISNGYQISL	EEARRVKFNT	IETTLTHSSG	PASYGGPRKA		
GI	QLLIIPQHYA	VVKKAQRE.G	CAYIAFKTNP	NSMVSHIAGK	SSIIFRALPD	VLANAYRISR	EEAQLKHNR	GDEFGAFTPI	QYKSYQDVYN		
Cons	QLLVPPQNFA	VVKKAG-E-G	FE-IAFKTN-	NA-IS-LAG-	-SV-RALP--	V-ANSY-ISR	EEARRLKFN-	-E-FL--P-Q	-----A		
G4		637									
LA											
C1											
C2											
Cr	DA.....										
GI	A..AESS										
Cons	-----										

Figure 9. Comparison of Aligned Amino Acid Sequences of Legumin-Like Subunits.

Designations and references for the subunits can be found in Table 3. VR1 and VR2 refer to variable regions 1 and 2, respectively. Other abbreviations were specified in Figure 5. Determined (G4 and LA, or predicted [others]) N-terminal residues of the acidic polypeptides are underlined. Conserved positions of introns in corresponding genes are indicated (>). Cons refers to consensus or preferred amino acids.

shows that these same cysteine residues are conserved in other legumin-like subunits as well (Figure 9, positions 55 and 85). The conservation of these other two cysteines suggests that they could be involved in an intrachain disulfide linkage. The results of Staswick et al. (1984) support this possibility. They reported the likelihood of another disulfide bond in G2 between cysteine 10 (Figure 5, position 36) and another cysteine within 71 residues.

The alignments shown in Figure 9 reveal three regions of variability in the legumin-like subunits. One occurs at the carboxyl ends of the acidic chains in the subunits, and is especially variable. This region has been described earlier (Argos et al., 1985), and is referred to as the hypervariable region (HVR). The second region of variability was first observed by Simon et al. (1985), and is referred to as variable region 1 or VR1. It is located between positions 133 to 178 in Figure 9. The VR1 region of cruciferin, a *Brassica* storage protein that lacks HVR sequences (Figure 9), has extended runs of glutamine and glycine that are not observed in the other legumin-like proteins. The position of the glutamine/glycine insertion in cruciferin corresponds to the cleavage position (Figure 5, position 124/125; Figure 9, position 140/141) in the glycinin G4 acidic chain (Staswick and Nielsen, 1983) and suggests that this part of the molecule is located on the surface of the molecule. It is possible that VR1 of cruciferin occupies the same spatial location in the folded protein as does the HVR. In support of this notion, the cysteine residues involved in the disulfide bond that link the acidic and basic chains of the glycinin subunits occur within 10 amino acids of each region. This suggests that these sequences are in close proximity in the folded protein. Finally, the third region of internal heterogeneity, variable region 2 (VR2), spans amino acid positions 245 to 286 of the alignments shown in Figure 9. The composition of VR2 also tends to be rich in glutamine, a characteristic of the other variable regions.

The three variable regions in the 11S proteins may be of importance in efforts to improve seed nutritional quality by protein engineering. The naturally occurring variability in the regions implies that they do not perform critical functions in assembly of the subunits into oligomers. Surface probability plots of the 11S subunits reveal a striking similarity between the variable regions and positions of high surface probability (data not shown). It is likely that the variable regions occur as loops at the surface of the polypeptides. The validity of this idea is supported by the observation that proteolytic cleavages occur at or near the variable regions in the subunits. These arguments suggest that changes engineered into the variable regions, and in particular the HVR, would be less likely to adversely affect assembly of the subunits than changes in areas where the molecules have been more highly conserved during evolution. Consistent with this prediction, deletions and insertions in the glycinin HVR have no detectable effect on G4 subunit self-assembly in vitro (Dickinson, 1988).

Glycinin Genes Are Coordinately Expressed

Figures 7 and 8 show that there is a striking correlation between the expression patterns of glycinin gene family members. Each glycinin gene is expressed during embryogenesis and is not detectably expressed in postgermination cotyledons or mature plant organ systems (Figure 7). In addition, each glycinin gene is expressed in the same temporal framework during embryogenesis (Figures 7 and 8). This differs significantly from the situation that occurs for the soybean β -conglycinin gene family. In that storage protein gene family, mRNAs encoding the α/α' and β subunits accumulate and decay at different times during seed development (Naito et al., 1988; J.J. Harada and R.B. Goldberg, unpublished results). Glycinin gene expression is regulated primarily at the transcriptional level (Figure 8; Goldberg et al., 1981a), as is the case for other seed protein genes (Walling et al., 1986; Goldberg et al., 1989). Figure 8 shows, however, that posttranscriptional events are also important in regulating seed mRNA levels.

Our results indicate that glycinin genes are activated transcriptionally at the same time during embryogenesis, are repressed coordinately prior to seed dormancy, and remain transcriptionally repressed during the rest of the life cycle (Figure 8), even though these genes can reside in different chromosomal domains (Figure 1; Cho et al., 1989a). This suggests that members of the glycinin gene family share common cis-control elements that are responsible for programming their coordinate expression during the soybean life cycle. Comparisons of the 5'-flanking regions of glycinin genes indicate that they are highly homologous (Figure 3 and Table 2), and this probably reflects the evolutionary relatedness of glycinin genes. However, consensus sequences that could play a role in glycinin gene expression have been identified in the 5' region of all glycinin genes. These include the 28-bp "legumin box" (Baumlein et al., 1986; Gatehouse et al., 1986), the 5'-CATGCATG-3' RY repeat element (Dickinson et al., 1988), and the CACA element (Goldberg, 1986). Both the RY and CACA elements have been found in the 5' region of other seed protein genes.

The precise DNA sequence elements responsible for regulating glycinin gene expression and the expression of other seed protein genes are not yet known (Goldberg et al., 1989). Recent transformation studies with the *Gy*₁ glycinin gene indicate that only 65 nucleotides 5' to the transcription start site are required for embryo-specific expression, and that a region from nucleotides -454 to -64 is required for correct quantitative expression levels (R.B. Goldberg, T.L. Sims, and J. Truettner, unpublished results). The quantitative region contains two 5'-CATGCATG-3' sequences (-109 and -253) and a CACA element (-427). DNA binding protein experiments showed that the CACA element interacts with an embryo nuclear protein that protects it from DNase I digestion (L. Perez-Grau and R.B. Goldberg, unpublished results). Clearly, the

DNA sequences and protein factors necessary for regulating members of the glycinin gene family, as well as the amino acid regions critical for glycinin protein assembly during seed development, remain to be determined.

METHODS

Plant Material

Seeds of Forrest and Dare soybean varieties were obtained from the U.S. Department of Agriculture soybean germplasm collection in Stoneville, MS. Embryos, leaves, roots, stems, and postgermination cotyledons were developmentally staged, frozen in liquid nitrogen, and stored at -80°C (Goldberg et al., 1981a, 1981b).

Cloning Vectors and Bacterial Strains

Escherichia coli K12 strain HB101 (Boyer and Roulland-Dussoix, 1969) was used to propagate plasmids pBR322 and pBR325 (Bolivar et al., 1977; Bolivar, 1978). Strain K802 (Wood, 1966) was used to grow the Charon 4 recombinant phage (Blattner et al., 1977). Strains JM103, JM105, and JM107 were used as hosts for M13-derived phage vectors.

Isolation and Labeling of Nucleic Acids

Soybean nuclear DNA, Charon 4 recombinant phage DNA, and plasmid DNAs were isolated using procedures described previously (Fischer and Goldberg, 1982; Marco et al., 1984). Phage and plasmid DNAs were labeled by nick-translation with ^{32}P -labeled deoxynucleotide triphosphates according to procedures specified by Bethesda Research Laboratories. Poly(A⁺) polysomal RNA was prepared from soybean leaves, embryos, roots, stems, and postgermination cotyledons by procedures described elsewhere (Kamalay and Goldberg, 1980; Goldberg et al., 1981a, 1981b; Tumer et al., 1981; Cox and Goldberg, 1988). Random-primed cDNA was synthesized according to the procedures of Van Ness and Hahn (1980). Nuclear RNA was labeled in vitro by isolating nuclei from leaves or embryos and allowing nascent transcripts to elongate in the presence of ^{32}P -UTP (Luthe and Quatrano, 1980a, 1980b; Walling et al., 1986; Cox and Goldberg, 1988). The labeled nuclear RNAs were purified according to procedures described previously (Walling et al., 1986; Cox and Goldberg, 1988), and hybridized with excess cloned soybean DNA sequences immobilized on nitrocellulose discs. The hybrids were washed under stringent conditions (0.36 M Na^+ , 65°C) and treated with RNase, and the counts per minute were measured as described by Walling et al. (1986). Fraction input disintegrations per minute (FDPM) were calculated using the formula:

$$\text{FDPM} = \frac{(\text{cpm/filter}) (\text{mRNA size})}{(\text{input cpm}) (\text{cDNA insert size}) (0.2)}$$

A 20% filter hybridization efficiency was calculated as described by Walling et al. (1986) and cpm/filter were corrected for nonspecific hybridization with pBR322 vector sequences.

Gel Blot Hybridization

The DNA gel blots were carried out as described by Southern (1975) and Wahl et al. (1979). For RNA gel blots, poly(A⁺) polysomal RNA was separated by electrophoresis in methylmercury hydroxide agarose gels (Bailey and Davidson, 1976), and analyzed using procedures described by Alwine et al. (1979) and by Thomas (1980). RNA dot blots were performed as described by Bletz et al. (1983).

Isolation and Mapping of Genomic and cDNA Clones

A library of genomic DNA sequences for the variety Dare was constructed by ligating 20-kb HaeIII and AluI restriction fragments into the Charon 4 λ cloning vector (K.D. Jofuku and R.B. Goldberg, unpublished results). A DNA library of the cultivar Forrest was prepared using random partial EcoRI fragments ligated into Charon 4 (Fischer and Goldberg, 1982). Genomic clones that contained the *Gy*₁, *Gy*₂, and *Gy*₃ were selected from the libraries using cDNA clone A-28 (Goldberg et al., 1981a). Genomic clones that contained the *Gy*₄ and *Gy*₅ glycinin genes were purified as described by Scallan et al. (1985). Specific phage DNA restriction fragments were recloned into plasmid vectors pBR322, pBR325, pUC8, or pUC12. The cDNA clones that hybridized specifically to individual glycinin genes (e.g. pG3-1) were obtained from a cDNA library of midmaturation stage embryo mRNA sequences (Goldberg et al., 1981a) by colony hybridization (1 M Na^+ , 50% formamide, 30°C) with a mixture of labeled *Gy*₁, *Gy*₂, and *Gy*₃ sequences. Isolation of pG258 and pG23 cDNA clones was described by Scallan et al. (1985).

R-Loop and Heteroduplex Analysis

R-loops were formed between cloned genomic DNA and embryo mRNA using RNA excess conditions as described by Kaback et al. (1979). R-loop formation between cloned genomic DNA and leaf mRNA was accomplished by using DNA excess conditions (Kaback et al., 1981). Heteroduplexes between recombinant phage DNAs were formed as described by Shen and Maniatis (1980). Electron microscopy used to visualize the molecules was performed using parlodian-coated grids as described by Kaback et al. (1979). Internal double- and single-stranded ϕX174 size standards were used to convert length measurements made by electron microscopy into kilobase sizes (Fischer and Goldberg, 1982).

DNA Sequence Analysis

Nucleotide sequence analysis was carried out using either the method of Maxam and Gilbert (1980) or the Sanger technique (1977). The boundaries of glycinin genes were determined by S1 protection using procedures described by Berk and Sharp (1977) and Nomura and Ray (1980).

Sequence Comparisons

DNA sequences were analyzed using either the programs of Staden (1982) and Kanehisa (1982) or the programs of the Uni-

versity of Wisconsin Genetics Computer Group (Devereux et al., 1984). Nucleotide and amino acid sequence alignments were carried out using the method of Needleman and Wunsch (1970), using gap weight and gap length values of 5.0 and 0.3, respectively, as modified by Gribskov and Burgess (1986).

ACKNOWLEDGMENTS

The experiments reported in this paper were the result of cooperative research between the U.S. Department of Agriculture/Agricultural Research Service, the Indiana Agricultural Experiment Station, and the Department of Biology, University of California, Los Angeles. The research was supported in part by grants from the U.S. Department of Agriculture Competitive Grants Program (N.C.N. and R.B.G.) and the American Soybean Association (N.C.N.). R.L.F. and T.L.S. were supported by National Institutes of Health Postdoctoral Fellowships, and G.N.D. was supported by a McKnight Foundation Predoctoral Fellowship. We thank Dr. John Harada for assistance with the *in vitro* transcription studies. This is Journal Paper No. 11,722 from the Indiana Agricultural Experiment Station.

Received January 20, 1989.

REFERENCES

- Alwine, J.C., Kemp, D.J., Parker, B.A., Reisner, J., Stark, G.R., and Wahl, G.M. (1979). Detection of specific RNAs or specific fragments of DNA by fractionation in gels and transfer to diazobenzoyloxymethyl paper. *Methods Enzymol.* **68**, 220–242.
- Argos, P., Narayana, S.V.L., and Nielsen, N.C. (1985). Structural similarity between legumin and vicilin storage proteins from legumes. *EMBO J.* **4**, 1111–1117.
- Badley, R.A., Atkinson, D., Hauser, H., Oldani, D., Green, J.P., and Stubbs, J.M. (1975). The structure, physical and chemical properties of the soybean protein glycinin. *Biochim. Biophys. Acta* **412**, 214–228.
- Bailey, J.M., and Davidson, N. (1976). Methylmercury as a reversible denaturing agent for agarose gel electrophoresis. *Anal. Biochem.* **70**, 75–85.
- Baumlein, H., Wobus, U., Pustell, J., and Kafatos, F.C. (1986). The legumin gene family: Structure of a B type gene of *Vicia faba* and a possible legumin gene specific regulatory element. *Nucl. Acids Res.* **14**, 2707–2720.
- Benoist, C., Ohare, K., Breathnach, R., and Chambon, P. (1980). Ovalbumin gene-sequence of putative control regions. *Nucl. Acids Res.* **8**, 127–142.
- Berk, A.J., and Sharp, P.A. (1977). Sizing and mapping of early adenovirus mRNAs by gel electrophoreses of S1 endonuclease-digested hybrids. *Cell* **12**, 721–732.
- Blattner, F.R., Williams, B.G., Blechl, A.E., Denniston-Thompson, K., Faber, H.E., Furlong, L.A., Grunwald, J.F., Keefer, D.O., Moore, D.D., Schumm, J.W., Sheldon, E.L., and Smithies, O. (1977). Charon phages: Safer derivatives of bacteriophage lambda for DNA cloning. *Science* **196**, 161–169.
- Bletz, G.A., Jacobs, K.A., Eickbush, T.H., Cherbos, P.T., and Kafatos, F.C. (1983). Isolation of multigene families and determination of homologies by filter hybridization methods. *Methods Enzymol.* **100**, 266–285.
- Bolivar, F. (1978). Construction and characterization of new cloning vehicles. III. Derivatives of plasmid pBR322 carrying unique EcoRI sites for selection of EcoRI generated recombinant DNA molecules. *Gene* **4**, 121–136.
- Bolivar, F., Rodriguez, R.L., Greene, P.J., Betlach, M.T., Heynecker, H.L., Boyer, H.W., Crosa, J.H., and Falkow, S. (1977). Construction and characterization of new cloning vehicles. II. A multi-purpose cloning system. *Gene* **2**, 95–136.
- Bornstein, P., and Balian, G. (1977). Cleavage at Asn-Gly bonds with hydroxylamine. *Methods Enzymol.* **47**, 132–145.
- Boyer, H.W., and Roulland-Dussoix, D. (1969). A complementation analysis of the restriction and modification of DNA in *Escherichia coli*. *J. Mol. Biol.* **41**, 459–472.
- Breathnach, R., Benoist, C., O'Hare, K., Gannon, F., and Chambon, P. (1978). Ovalbumin gene: Evidence for a leader sequence in mRNA and DNA sequences at the exon-intron boundaries. *Proc. Natl. Acad. Sci. USA*, **75**, 4853–4857.
- Chlan, C.A., Pyle, J.B., Legocki, A.B., and Dure, L. (1986). Sequences and genomic organization of the α -globulin (vicilin) genes of cottonseed. *Plant Mol. Biol.* **7**, 475–489.
- Cho, T.-J., Davies, C.S., and Nielsen, N.C. (1989a). Inheritance and organization of glycinin genes in soybean. *Plant Cell* **1**, 329–337.
- Cho, T.-J., Davies, C.S., Fischer, R.L., Tumer, N.E., Goldberg, R.B., and Nielsen, N.C. (1989). Molecular characterization of an aberrant allele for the Gy₃ glycinin gene: A chromosomal rearrangement. *Plant Cell* **1**, 339–350.
- Chrispeels, M.J., Higgins, J.V., and Spencer, D. (1982). Assembly of storage protein oligomers in the endoplasmic reticulum and processing of the polypeptides in the protein bodies of developing pea cotyledons. *J. Cell Biol.* **93**, 306–313.
- Cox, K.H., and Goldberg, R.B. (1988). Analysis of gene expression. In *Plant Molecular Biology: A Practical Approach*, C.H. Shaw, ed (Oxford: IRL Press), pp. 1–35.
- Derbyshire, E., Wright, D.J., and Boulter, D. (1976). Legumin and vicilin, storage proteins of legume seeds. *Phytochemistry* **15**, 3–24.
- Devereux, J., Haeberli, P., and Smithies, O. (1984). A comprehensive test of sequence analysis for the VAX. *Nucl. Acids Res.* **12**, 387–395.
- Dickinson, C.D. (1988). Assembly properties of glycinin subunits: Development of a novel *in vitro* assembly assay. PhD Thesis, Purdue University, West Lafayette, IN.
- Dickinson, C.D., Evans, R.P., and Nielsen, N.C. (1988). RY-repeats are conserved in the 5'-flanking regions of legume seed-protein genes. *Nucl. Acids Res.* **16**, 371.
- Domoney, C., and Casey, R. (1985). Measurement of gene number for seed storage proteins in *Pisum*. *Nucl. Acids Res.* **13**, 687–699.
- Fischer, R.L., and Goldberg, R.B. (1982). Structure and flanking regions of soybean seed protein genes. *Cell* **29**, 651–660.
- Fitzgerald, M., and Shenk, T. (1981). The sequence 5'-AAUAAA-

- 3' forms part of the recognition site for polyadenylation of late SV40 mRNAs. *Cell* **24**, 251–260.
- Fryberg, E.A., Kindle, K.L., Davidson, N., and Sodja, A.** (1980). The actin genes of *Drosophila*: A dispersed multigene family. *Cell* **19**, 365–378.
- Fukazawa, C., Momma, T., Hirano, H., Harada, K., and Udaka, K.** (1985). Glycinin A₃B₄ mRNA. *J. Biol. Chem.* **260**, 6234–6239.
- Gatehouse, J.A., Evans, I.M., Croy, R.R.D., and Boulter, D.** (1986). Differential expression of genes during legume seed development. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **314**, 367–384.
- Goldberg, R.B.** (1986). Regulation of plant gene expression. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **314**, 343–353.
- Goldberg, R.B., Hoscheck, G., Ditta, G.S., and Breidenbach, R.W.** (1981a). Developmental regulation of clone superabundant embryo mRNAs in soybean. *Dev. Biol.* **83**, 218–231.
- Goldberg, R.B., Hoscheck, G., Tam, S.H., Ditta, G.S., and Breidenbach, R.W.** (1981b). Abundance, diversity and regulation of mRNA sequence sets in soybean embryogenesis. *Dev. Biol.* **83**, 201–217.
- Goldberg, R.B., Barker, S.J., and Perez-Grau, L.** (1989). Regulation of gene expression during plant embryogenesis. *Cell* **56**, 149–160.
- Gribskov, M., and Burgess, R.R.** (1986). Sigma factors from *E. coli*, *B. subtilis*, phage T4 are homologous proteins. *Nucl. Acids Res.* **67**, 6745–6763.
- Kaback, D.B., Angerer, L.M., and Davidson, N.** (1979). Improved methods for the formation and stabilization of R-loops. *Nucl. Acids Res.* **6**, 2499–2517.
- Kaback, D.G., Rosbash, M., and Davidson, N.** (1981). Determination of cellular RNA concentrations by electron microscopy of R-loop containing DNA. *Proc. Natl. Acad. Sci. USA* **78**, 2820–2824.
- Kamalay, J.C., and Goldberg, R.B.** (1980). Regulation of structural gene expression in tobacco. *Cell* **19**, 935–946.
- Kanehisa, M.I.** (1982). Los Alamos sequence analyses package for nucleic acids and proteins. *Nucl. Acids Res.* **10**, 183–196.
- Kitamura, K., Davies, C.S., and Nielsen, N.C.** (1984). Inheritance of alleles for *Cgy*, and *Gy*₄ storage protein genes in soybean. *Theor. Appl. Genet.* **68**, 253–257.
- Larkins, B.A.** (1981). Seed storage proteins: Characterization and biosynthesis. In *The Biochemistry of Plants: A Comprehensive Treatise*, Vol. 6, P.K. Stumpf, and E.E. Conn, eds (New York: Academic Press), pp. 449–489.
- Lee, J.S., and Verma, D.P.S.** (1984). Structure and chromosomal arrangement of leghemoglobin genes in kidney bean suggest divergence in soybean leghemoglobin gene loci following tetraploidization. *EMBO J.* **3**, 2745–2752.
- Lei, M.G., Tyrell, D., Bassette, R., and Reeck, G.R.** (1983). Two-dimensional electrophoretic analysis of soybean proteins. *J. Agric. Food Chem.* **31**, 963–968.
- Lutcke, H.A., Chow, K.C., Mickel, F.S., Moss, K.A., Kern, H.F., and Scheele, G.A.** (1987). Selection of AUG differs in plants and animals. *EMBO J.* **6**, 43–48.
- Luthe, D.S., and Quatrano, R.S.** (1980a). Transcription in isolated wheat nuclei. I. Isolation of nuclei and elimination of endogenous ribonuclease activity. *Plant Physiol.* **65**, 305–308.
- Luthe, D.S., and Quatrano, R.S.** (1980b). Transcription in isolated wheat nuclei. II. Characterization of RNA synthesized in vitro. *Plant Physiol.* **65**, 309–313.
- Lycett, G.W., Croy, R.R.D., Shirsat, A.H., and Boulter, D.** (1984). The complete nucleotide sequence of a legumin gene from pea (*Pisum sativum* L.). *Nucl. Acids Res.* **12**, 4493–4506.
- Marco, Y.A., Thanh, V.H., Tumer, N.E., Scallion, B.J., and Nielsen, N.C.** (1984). Cloning and structural analysis of DNA encoding an A₂/B_{1a} subunit of glycinin. *J. Biol. Chem.* **259**, 13436–13441.
- Matta, N.K., Gatehouse, J.A., and Boulter, D.** (1981). Molecular and subunit heterogeneity of legumin of *Pisum sativum* L. (Garden pea)—A multidimensional gel electrophoretic study. *J. Exp. Bot.* **32**, 1295–1307.
- Maxam, A.M., and Gilbert, W.** (1980). Sequencing end-labeled DNA with base-specific chemical cleavages. *Methods Enzymol.* **65**, 499–560.
- Meinke, D.W., Chen, J., and Beachy, R.N.** (1981). Expression of storage protein genes during soybean seed development. *Planta* **153**, 130–139.
- Miller, A.** (1975). Biochemistry of legume seed proteins. *Annu. Rev. Plant Physiol.* **26**, 53–72.
- Moreira, M.A., Hermodson, M.A., Larkins, B.A., and Nielsen, N.C.** (1979). Partial characterization of the acidic and basic polypeptides of glycinin. *J. Biol. Chem.* **254**, 9921–9926.
- Moreira, M.A., Hermodson, M.A., Larkins, B.A., and Nielsen, N.C.** (1981). Comparison of the antigenic properties of the glycinin polypeptides. *Arch. Biochem. Biophys.* **210**, 633–642.
- Naito, S., Dubé, P.H., and Beachy, R.N.** (1988). Differential expression of conglycinin α' and β subunits in transgenic plants. *Plant Mol. Biol.* **11**, 109–123.
- Needleman, S.B., and Wunsch, C.D.** (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48**, 443–453.
- Nielsen, N.C.** (1984). The chemistry of legume storage proteins. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **304**, 287–296.
- Nielsen, N.C.** (1985). The structure and complexity of the 11S polypeptides in soybeans. *J. Am. Oil Chem. Soc.*, **62**, 1680–1686.
- Nielsen, N.C.** (1989). In vitro modification and assembly of soybean glycinin. In *Proceedings of the World Congress and Expo on Vegetable Protein Utilization in Human Foods and Animal Feedstuffs*, G. Willhite and K. Beery, eds (Urbana, IL: American Oil Chemists' Society), in press.
- Nomura, N., and Ray, D.S.** (1980). Expression of a DNA strand initiation sequence of ColE1 plasmid in a single-stranded DNA phage. *Proc. Natl. Acad. Sci. USA* **77**, 6566–6570.
- Okamuro, J.K., Jofuku, K.D., and Goldberg, R.B.** (1986). Soybean seed lectin gene and flanking nonseed protein genes are developmentally regulated in transformed tobacco plants. *Proc. Natl. Acad. Sci. USA* **83**, 8240–8244.
- Palmer, R.G.** (1978). Chromosome transmission and morphology of three primary trisomics in soybean (*Glycine max*). *Can. J. Gen. Genet. Cytol.* **18**, 131–140.
- Sanger, F., Nicklen, S., and Coulson, A.R.** (1977). DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. USA* **74**, 5463–5467.
- Scallion, B., Thanh, V.H., Floener, L.A., and Nielsen, N.C.** (1985). Identification and characterization of DNA clones encoding

- group-II glycinin subunits. *Theor. Appl. Genet.* **70**, 510–519.
- Shen, C.-K.J., and Maniatis, T.** (1980). The organization of repetitive sequences in a cluster of rabbit β -like globin genes. *Cell* **19**, 379–391.
- Shotwell, M.A., Afonso, C., Davies, E., Chesnut, R.S., and Larkins, B.A.** (1988). Molecular characterization of oat seed globulin. *Plant Physiol.* **87**, 698–704.
- Simon, A.E., Tenbarger, K.M., Scofield, S.R., Finkelstein, R.R., and Crouch, M.L.** (1985). Nucleotide sequence of a cDNA clone from *Brassica napus* 12S storage protein shows homology with legumin from *Pisum sativum*. *Plant Mol. Biol.* **5**, 191–201.
- Slightom, J.L., Blechl, A.E., and Smithies, O.** (1980). Human fetal G γ - and A γ -globin genes: Complete nucleotide sequences suggest that DNA can be exchanged between these duplicated genes. *Cell* **21**, 627–638.
- Southern, E.M.** (1975). Detection of specific DNA sequences among DNA fragments separated by gel electrophoresis. *J. Mol. Biol.* **98**, 503–517.
- Spencer, D.** (1984). The physiological role of storage proteins in seeds. *Philos. Trans. R. Soc. Lond B Biol. Sci.* **304**, 275–285.
- Staden, R.** (1982). An interactive graphics program for comparing and aligning nucleic acid and amino acid sequences. *Nucl. Acids Res.* **10**, 2951–2961.
- Staswick, P.E., and Nielsen, N.C.** (1983). Characterization of a soybean cultivar lacking certain glycinin subunits. *Arch. Biochem. Biophys.* **223**, 1–8.
- Staswick, P.E., Hermodson, M.A., and Nielsen, N.C.** (1981). Identification of the acidic and basic subunit complexes of glycinin. *J. Biol. Chem.* **256**, 8752–8755.
- Staswick, P.E., Hermodson, M.A., and Nielsen, N.C.** (1984). The amino acid sequence of the A₂B_{1a} subunit of glycinin. *J. Biol. Chem.* **259**, 13431–13435.
- Takaiwa, F., Kikuchi, S., and Oono, K.** (1986). The structure of rice storage protein glutelin precursor deduced from cDNA. *FEBS Lett.* **206**, 33–35.
- Takaiwa, F., Ebinuma, H., Kikuchi, S., and Oono, K.** (1987). Nucleotide sequence of a rice glutelin gene. *FEBS Lett.* **221**, 43–47.
- Thomas, P.** (1980). Hybridization of denatured RNA and small DNA fragments transferred to nitrocellulose. *Proc. Natl. Acad. Sci. USA* **77**, 5201–5205.
- Tumer, N.E., Thanh, V.H., and Nielsen, N.C.** (1981). Purification and characterization of mRNA from soybean seeds. *J. Biol. Chem.* **256**, 8756–8760.
- Tumer, N., Richter, J.D., and Nielsen, N.C.** (1982). Structural characterization of the glycinin precursors. *J. Biol. Chem.* **257**, 4016–4018.
- Van Ness, J., and Hahn, W.E.** (1980). Sequence complexity of cDNA transcribed from diverse mRNA population. *Nucl. Acids Res.* **18**, 4259–4269.
- Walling, L., Drews, G.M., and Goldberg, R.B.** (1986). Transcriptional and post-transcriptional regulation of soybean seed protein mRNA levels. *Proc. Natl. Acad. Sci. USA* **83**, 2123–2127.
- Wahl, G.M., Stern, M., and Stark, G.R.** (1979). Efficient transfer of large DNA fragments from agarose gels to diazobenzoylmethyl paper and rapid hybridization using dextran sulfate. *Proc. Natl. Acad. Sci. USA* **76**, 3683–3687.
- Wolf, W.J.** (1976). Chemistry and technology of soybeans. *Adv. Cereal Sci. Technol.* **11**, 325–377.
- Wood, W.B.** (1966). Host specificity of DNA produced by *Escherichia coli*: Bacterial mutations affected the restriction and modification of DNA. *J. Mol. Biol.* **16**, 325–377.

Characterization of the Glycinin Gene Family in Soybean

N. C. Nielsen, C. D. Dickinson, T. J. Cho, V. H. Thanh, B. J. Scallan, R. L. Fischer, T. L. Sims, G. N. Drews and R. B. Goldberg
Plant Cell 1989;1;313-328
DOI 10.1105/tpc.1.3.313

This information is current as of April 14, 2012

Permissions	https://www.copyright.com/ccc/openurl.do?sid=pd_hw1532298X&issn=1532298X&WT.mc_id=pd_hw1532298X
eTOCs	Sign up for eTOCs at: http://www.plantcell.org/cgi/alerts/ctmain
CiteTrack Alerts	Sign up for CiteTrack Alerts at: http://www.plantcell.org/cgi/alerts/ctmain
Subscription Information	Subscription Information for <i>The Plant Cell</i> and <i>Plant Physiology</i> is available at: http://www.aspb.org/publications/subscriptions.cfm