Plant Molecular Biology 7: 475–489, 1986

Developmental biochemistry of cottonseed embryogenesis and germination XVIII cDNA and amino acid sequences of members of the storage protein families

Caryl A. Chlan,[1] J. B. Pyle[1], A. B. Legocki[2] & Leon Dure III[1,3]
[1]*Department of Biochemistry, University of Georgia, Athens, GA 30602, U.S.A.*
[2]*Department of Biochemistry, University of Agriculture, Wolynska 35, Pozan, Poland*

Keywords: cotton storage proteins, cDNA sequence, protein sequence, codon usage, dot matrix sequence analysis, hydropathy analysis

Summary

We have sequenced cDNA clones representing each of the three distinct groups of storage proteins of the cotton seed. Characteristics of their mRNAs and derived proteins are given. Dot matrix analysis of the nucleotide and amino acid sequences shows that 2 of these groups of proteins have a great deal of vestigial homology at low stringency and should be considered subfamilies of a single storage protein gene family. The remaining group is quite distinct and should be considered a separate multigene family. It also can be divided into 2 subfamilies based on the presence or absence of glycosyl residues and other sequence differences.

These proteins are processed to smaller species during embryogenesis, and all of the mature storage proteins of cotton can be traced back to these 2 gene families.

In view of these relationships we propose that these 2 families be called the α and β globulins of cotton storage proteins, each comprised of an A and B subfamily.

Introduction

We have shown that the seed storage proteins of the cotton (Gossypium hirsutum) emanate from 2 sets of preproproteins of apparent size of 69 and 60 kD (1, 2). A single cDNA clone prepared from mRNA of embryonic cotyledons will arrest the translation of all the 69 kD preproproteins, whereas 2 different cDNA clones are required to arrest the translation of the 60 kD preproproteins (2). From this it would appear that 3 sets of genes give rise to the storage proteins of these species. However, the 2 sets of cDNA that each partially arrest the translation of the 60 kD preproproteins exhibit weak cross hybridization at very low criteria suggesting some vestigial homology between these cDNAs and thus their genes.

Sequence analysis of cDNA's representing each of the 3 sets and dot matrix and hydropathy analyses of these sequences reveal that the 2 sets of cDNAs encoding the 60 kD family of preproproteins are members of a single gene family.

Characteristics of the mRNAs derived from the cDNAs, codon usage, amino acid compositions and the putative cleavage sites for producing the mature storage proteins species also are presented here.

Materials and methods

Cotton plants from which mRNA was extracted for cDNA synthesis was *Gossypium hirsutum*, variety Coker 201. The preparation of cDNAs and their identification as storage protein representatives by hybrid selection and arrest have been described (2).

The nucleotides sequences were determined according to the chemical modification protocols of Maxam and Gilbert (3). DNA fragments after sub-

[3] To whom correspondence should be directed: (404) 542-2086 or Telex: 810-754-3908.

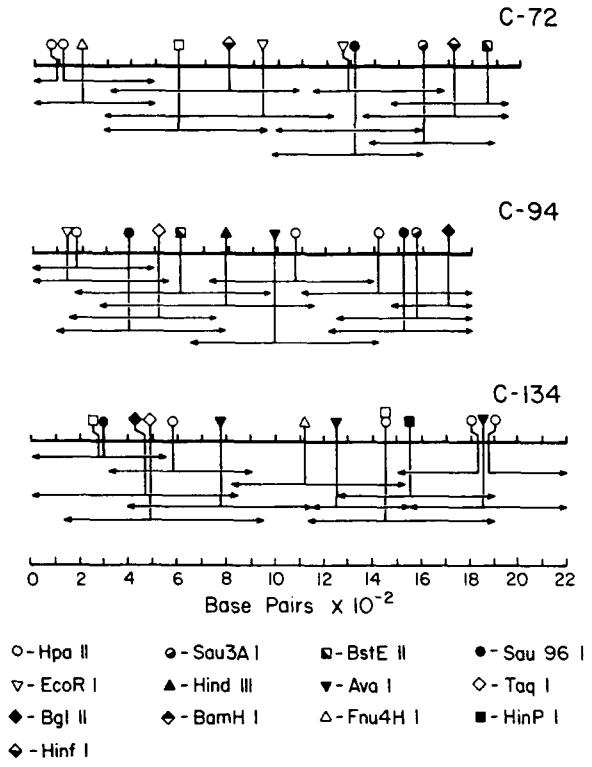

Fig. 1. Sequencing strategy. Clones were cut with the restriction endonucleases shown, labelled by site filling and sequencing in both directions from the cut site.

cloning were digested with restriction enzymes that leave recessed 3′ ends. The fragments were labelled by filling in the recessed ends with the Klenow fragment of DNA polymerase I using 32 P-labelled dNTPs. The end-labelled fragments were digested with a second restriction endonuclease and sequenced. The restriction endonucleases used and the length of sequence obtained from each site are given in Fig. 1. Sequencing was carried out in both directions from the cut sites, which provided the sequence of both DNA strands in over 90% of the cDNA spans.

To increase resolution, one side of the sequencing gels was covalently bound to one of the glass plates by the method of Garoff and Ansorge (4). After electrophoresis the gels were dried and autoradiographed.

The computer program for comparing nucleotide and amino acid sequences by dot matrix analyses was written by D. J. Neigel (UCLA, Los Angeles). This program allows for variation in the span of nucleotides and amino acids covered and in the stringency of matches per span.

The values for amino acid hydrophobicity used in the hydropathy analyses were those of Kyte and Doolittle (5).

Results

Characteristics of the cDNA clones

The notation used for identifying the cDNA clones sequenced is:

Clone C-72 – arrests the translation of the 69 kD preproproteins
Clone C-94 – arrests the translation of some of the 60 kD preproproteins
Clone C-134 – arrests the translation of the 60 kD preproproteins not arrested by C-94

Table 1. Characteristics of cDNA clone sequences and of the derived proteins. Protein sizes given in kDaltons.

	Preproprotein	Proprotein	Mature proteins[1]	N terminal Met	#AA in leader[2]	AATAAA	Poly A	Coding NT	3′ NT[3]	#AA[4]
C-72	69725[5]	67032	46464 20585	met-val missing[5]	23	multiple	–	1764	201	586
C-94	58212	56148	24400 20900 10500	met and 1 or 2 AA missing	19	+	+	1521	243	507
C-134	58713	56352	20900 20670 14900	+	22	multiple overlapping	–	1548	138	516

[1] From gel electrophoresis and sequence assuming cleavages given in text.
[2] Number of AA in cDNA clones.
[3] 3′ untranslated nucleotides contained in cDNA clones.
[4] The number of amino acids of preproprotein encoded by the cDNA clones. For complete preproprotein, add 2 AA to C-72 and 2 – 3 AA to C-94.
[5] Sequence of genomic DNA reveals that met-val is missing from the N terminal of C-72 preproprotein (ref. 16). These 2 AA are included in determination of size of this preproprotein.

Significant features of the nucleotide sequence are given in Table 1 and the sequences themselves in Figs. 2–4. Since the cDNAs were cloned with *Pst I*, the G and C tails are shown as part of the cDNA inserts in the Figures. Only the non-coding DNA strand is shown to represent the mRNA. All three clones have one or more AATAAA poly (A) addition sites in the 3 untranslated region, although

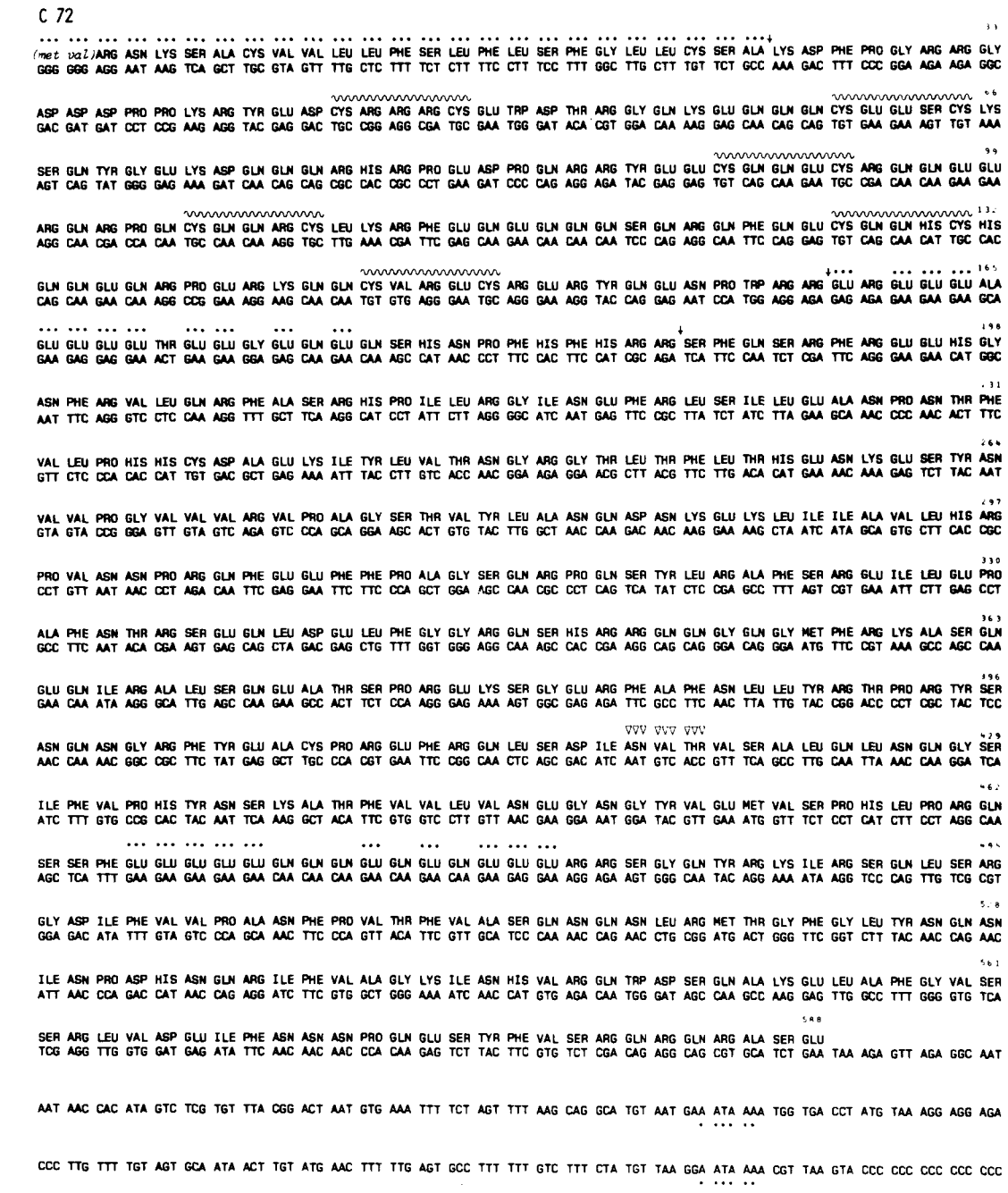

Fig. 2. Nucleotide and amino acid sequence of cDNA C-72. N terminal leader sequence is over-dotted; CysxxxCys sequences are indicated by wavy overlines; clusters of Glu residues are over-dotted; glycosylation site is disignated by ▽; poly A signal sites are under-dotted. Arrows indicate possible processing cleavage sites.

the sequence AATATA may carry out this function in C-94 since it is within 25 nucleotides of the poly (A) tail (marked on the Figures). The other two clones do not have poly (A) tails. Palindromic tracts capable of forming hairpin structures are found in the 3′ untranslated regions of C-94 and C-134 (marked on the Figures).

A curious anomaly was found in clone C-134; namely, the coding sequence from nucleotide 1344 to nucleotide 1728 is found in the inverse complement orientation beginning with the 1st 5′ cDNA nucleotide (not counting the poly G of the cDNA construct) and ending 2 nucleotides upstream from the initiating Met codon (denoted in Fig. 4). Only

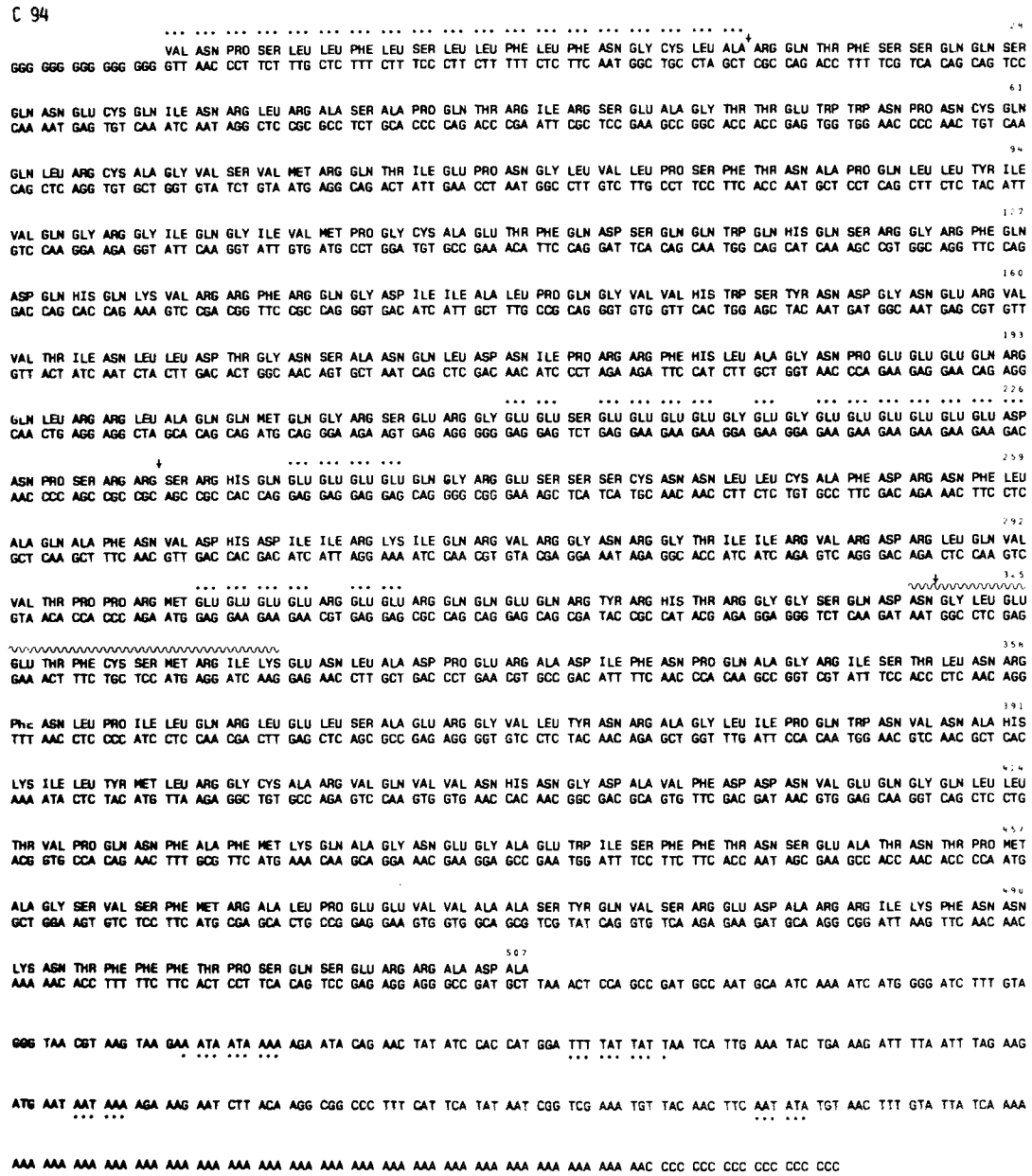

Fig. 3. Nucleotide and amino sequence of cDNA C-94. Wavy overline indicates longest stretch of homology with C-134. Other symbols as in Figure 2.

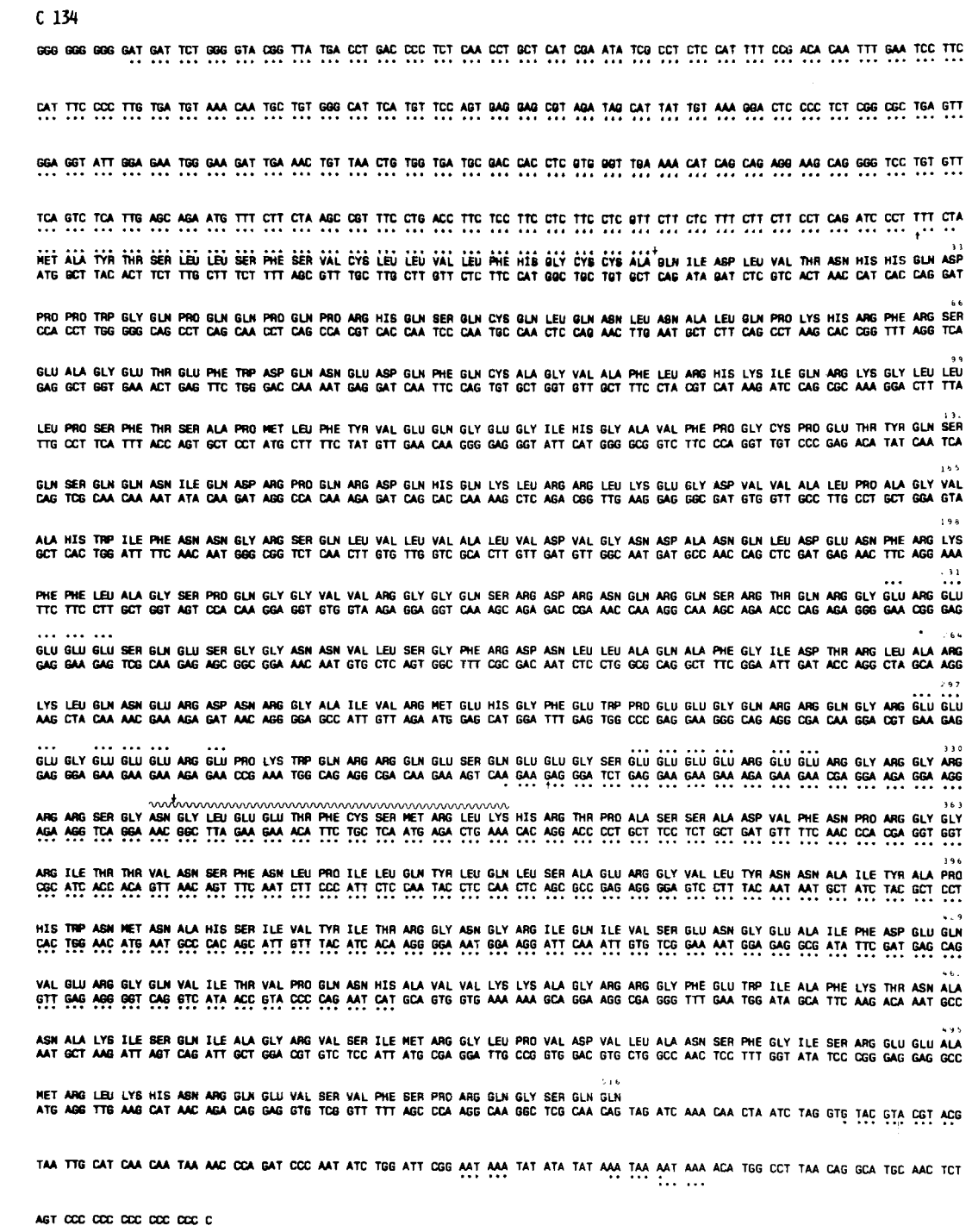

Fig. 4. Nucleotide and amino acid sequence of cDNA C-134. Under-dotted regions show inverse complement artifact as well as poly A signal sites. Other symbols as in Figure 2.

a single mismatch is present in this 385 nucleotide sector which is 6 nucleotides upstream from the Met codon. Obviously this sequence, which would create a very large hairpin in the mRNA itself, cannot exist as a functional molecule in the cell, and, thus, must be an artifact generated in cDNA synthesis. We mention this because we have observed the same phenomenon in 2 other cDNA prepared from cotton mRNA. Thus it is not an isolated incident. An explanation for how a 3′ sequence may end up in the inverse complement orientation at the 5′ end of a cDNA is presented in Fig. 5.

This proposed mechanism posits a fortuitous hybridization of the 3′ end of a nearly full length cDNA first strand with a short complementary region internal to the 3′ end. Second strand synthesis is postulated to add on to this loop structure making the complement of the 1st strand. This is what is considered to happen in normal double stranded cDNA synthesis. What is new in this model is the formation of a second loop causing continued cDNA second strand synthesis to form the complement of the portion of the 2nd strand already made. Continued 2nd synthesis gives a double stranded cDNA longer than the cognate mRNA and containing at the 5′ end (mRNA) in the inverse orientation the complement of the 3′ region.

This idea is further strengthened by the fact that the fortuitous complementarity that forms the 1st loop, being fortuitous, may not be perfect. In fact in clone C-134 there is a T/C mismatch 5 nucleotides from the beginning of the inverse complement region. There are no mismatches thereafter (379 nucleotides) since what is formed beyond the hybridized region is the result of 2nd strand synthesis. Others have noted similar cloning errors near the 5′ end as well (7, 8).

The sequence downstream from the artifact has been verified by the sequencing of genomic DNA (data not given).

Characteristics of derived proteins

Gross characteristics of the derived proteins are given in Table 1 and their sequences in Figs. 2–4.

The N-terminal Met is found on only 1 of the 3

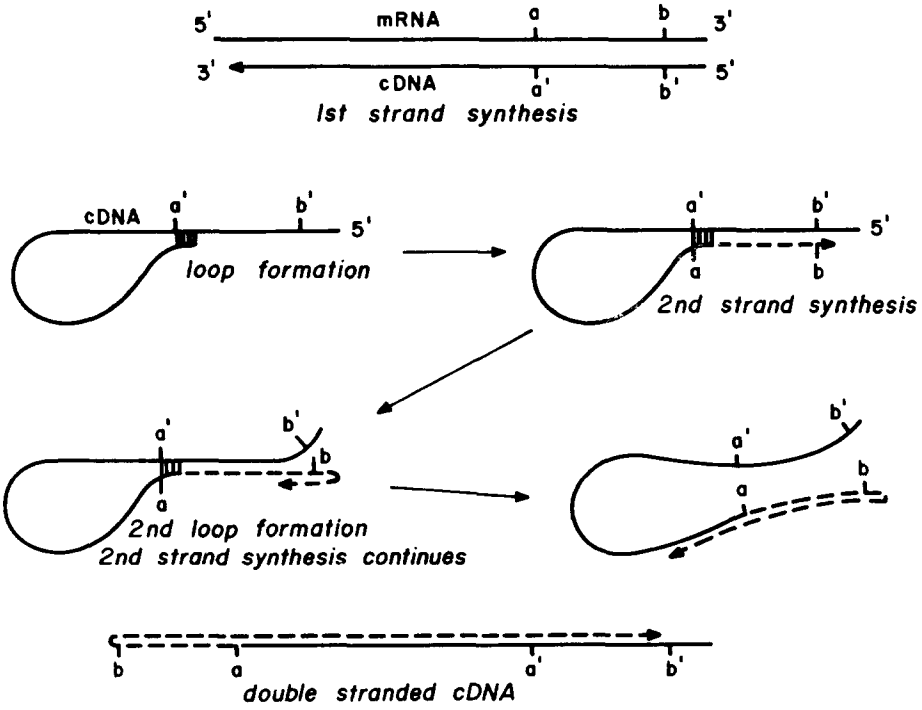

Fig. 5. Model attempting to explain origin of the cloning artifact found in cDNA C-134 (and in other non-storage protein cDNAs). Model is explained in the text.

cDNA clones (C-134). However, the other clones are very close to including the N-terminal coding region. The 5′ leader sequence is quite obvious in the sequence of all these proteins (see also Fig. 8, the hydropathy analysis of the proteins). Furthermore, a genomic sequence for the protein family typified by C-72 (16) shows that only a Met-Val is lacking from the coding region in this clone. Although the N terminal amino acids are blocked in the mature proteins, the cleavage point for the leader sequence is also unmistakable in all 3 clones using the formulation of von Heijnes (9). The leader sequences are designated on the Figures.

C-72 protein

The C-72 clone arrests the synthesis of the family of preproproteins of 69 kD which gives rise, after processing, to the mature proteins of 46.5 and 52 (glycosylated) kD and smaller fragments of about 20.5 kD (1). The precise endoproteolytic site has not been determined. However, measurements of molecular weights and calculated isoelectric points deduced from amino acid sequence, when compared with 2D gel profiles of the mature proteins, allow for a gross placement of this site (1) as does the amino acid composition of the 46.5 and 52 kD proteins determined some years ago (23). Figure 2 designates 2 possible sites (designated by arrows) that follow Arg-Arg pairs. Contiguous pairs of basic residues are common endoproteolytic cleavage sites (10), and, in the case of the specific protein corresponding to C-72, the Arg Arg/Ser site is a possible site since it yields 2 fragments with the molecular weights and pIs observed on 2D gels (1). Since the N terminal amino acid of the 46.5 and 52 kD larger fragment resulting from this cleavage is blocked (11), this cleavage should leave an amino acid suitable for substitution. Either the Glu or Ser residues shown satisfy this requirement. Of course other points of cleavage in this region of the proprotein may be proposed, but we have based the deduced amino acid composition of the mature proteins (Table 2) on the cleavage site given above. The C-72 protein has regions of concentrated Glu

Table 2. Representative amino acid composition of the cotton storage protein families.

	α Globulin family (cDNA C-72)				β Globulin family Subfamily A (cDNA C-94)		Subfamily B (cDNA C-134)	
	Preproprotein*	Proprotein	Mature protein: large fragment	Cys-rich fragment	Preproprotein**	Proprotein	Preproprotein	Proprotein
ASP	16	16	8	8	19	19	17	17
ASN	34	33	31	2	39	37	31	31
GLU	68	68	36	32	49	49	47	47
GLN	69	69	35	34	47	47	50	50
ARG	68	67	40	27	54	54	52	52
LYS	18	17	10	7	7	7	14	14
HIS	17	17	11	6	9	9	15	14
SER	41	37	33	4	32	30	33	30
THR	15	15	13	2	20	20	14	13
CYS	16	14	2	12	9	8	7	4
PRO	29	29	20	9	21	20	22	22
GLY	28	27	22	5	33	32	46	45
ALA	25	23	22	1	33	32	33	31
VAL	33	30	29	1	28	27	32	30
ILE	16	16	16	0	23	23	21	21
LEU	35	29	28	1	38	31	37	32
PHE	37	34	29	5	25	22	24	22
TYR	16	16	12	4	6	6	7	6
TRP	3	3	1	2	6	6	7	7
MET	4	3	3	0	9	9	7	6
	588	563	401	162	507	408	516	494
Mol wt	69725	67032	46464	20585	58212	56148	58713	56352

* The N terminal MetVal, missing from the cDNA sequence but known from a genomic sequence, has been added in this calculation.
** This clone lacks 1 – 3 amino acids of the N terminal. Thus these values are 1 – 3 amino acids low.

residues as do the other two storage proteins which are designated on the Figures. These concentrations of negative charge are thought to explain the 'negative staining' characteristics of these storage proteins (12) and help in identifying the cleavage fragments on 2D gels (1). C-72 protein large fragment contains a glycosylation site (designated in Fig. 2) and thus must represent a protein of the 52 kD group of this family of storage proteins. Other sequenced clones of this family lack this site and must represent members of the 46.5 kD group (16).

The most curious feature of the amino acid composition of this family of proteins is the composition of the mature small fragment (~20 kD). This polypeptide is very highly charged and contains 12 Cys residues arranged as 6 CysxxxCys sectors. The amino acid composition is unlike any reported for a plant storage protein and more closely resembles the Cys-rich domains of mammalian cell surface receptor proteins (13, 14, 15). It would seem that this domain has been added to this storage protein family from elsewhere as discussed in ref 16. The sequence of the large fragment of the C-72 protein shares homology with vicilin proteins of legumes (6).

C-94 and C-134 proteins

These proteins superficially do not seem related in sequence. They do have in common regions of concentrated Glu residues and a 13 amino acid region that almost certainly comprises a processing cleavage site for generating some of the mature cotton proteins (marked on Figs. 3, 4). Homologous sequences are found in the legumins of legumes (17, 18, 19) the cruciferins of Brassica (20) and the globulins of Avena (21). This sequence is known to contain the cleavage site Asn/Gly for the processing of the legumins (17). Thus the conservation of this sequence must represent the conservation of an endoprotease recognition site. This feature and others show a common heritage for these cotton proteins, the legumins family of legumes, the cruciferins of Brassica and the oat globulins (6, 22).

Amino acid composition and codon usage

Table 2 gives the amino acid compositions and molecular weights of the preproproteins, proproteins and mature processed proteins encoded in the representive cDNAs. It should be remembered that these cDNAs represent only one of several genes comprising the 2 small multigene families.

The amino acid compositions of the mature protein species are not given for C-94 and C-134, since all the cleavage sites have not been demonstrated. Both proteins are undoubtedly cleaved twice (not counting the removal of the leader sequences). The first cleavage quite likely occurs between the Asn/Gly of the tract of amino acids conserved in many angiosperm storage proteins as pointed out above. Cleavage here generates a small fragment that in each case has a molecular weight and pI of observed mature storage proteins (see ref 1 for other considerations). The large fragment in the case of C-94 has a transient existance in embryogenesis and can be seen on 2D gels (1). Its second cleavage is shown in Fig. 3 to occur in a ArgArg/Ser tract shown in the Figure since this would produce fragments with the characteristics of mature protein species (1). The second cleavage in C-134 can not be speculated at this point, although the mature proteins derived from this cleavage can be identified on 2D gels (1). The identification of the products of the second cleavage of these proteins is borne out by the disulfide linkages observed between mature species (22).

Codon usage in these three sequences is given in Table 3. Noteworthy here is the infrequent use of XCG. The summation of all three sequences shows that in the case of Ser where TCG is one of 6 code words, it is used on by 9/106 times; in the case Pro, Thr and Ala, all having 4 code words, XCG is used only 8/72, 4/49 and 9/91 times respectively.

Relationship among sequences

In order to search for homologous regions of sequence that may suggest evolutionary relations between these proteins, the nucleotide and amino acid sequences were compared by dot matrix analysis. In Fig. 6 the nucleotide sequences are compared at 2 levels of stringencies. In the left panels nucleotides in groups of 4 (1–4, 2–5, etc.) are compared and only a perfect match yields a dot. Using this low span of nucleotides, in which the random probability of a match is 1 in 256 sets of 4, considerable background is generated.

In the right panel a span of 12 nucleotides is used and 9 matches in position are required to yield a

Table 3. Codon usage.

		Preproproteins:			Proproteins:			Summation: (preproproteins)
		C-72	C-94	C-134	C-72	C-94	C-134	
Phe	TTT	10	6	8	8	4	7	24
	TTC	27	19	16	26	18	15	62
Leu	TTA	4	1	2	4	1	2	7
	TTG	11	4	9	9	3	7	24
	CTT	11	10	10	8	7	8	31
	CTC	5	17	10	4	15	9	32
	CTA	2	3	3	2	2	3	8
	CTG	2	3	3	2	3	3	8
Ile	ATT	4	12	11	4	12	11	27
	ATC	7	10	4	7	10	4	21
	ATA	5	1	6	5	1	6	12
Val	GTT	10	5	13	9	4	11	28
	GTC	7	9	6	7	9	6	22
	GTA	5	4	3	4	4	3	12
	GTG	10	10	10	10	10	10	30
Ser	TCT	10	5	5	8	4	3	20
	TCC	5	9	5	4	8	5	19
	TCA	9	6	5	8	6	5	20
	TCG	2	2	5	2	2	5	9
	AGT	6	3	6	6	3	6	15
	AGC	9	7	7	9	7	6	23
Pro	CCT	11	8	9	11	7	9	28
	CCC	3	5	4	3	5	4	12
	CCA	11	6	7	11	6	7	24
	CCG	4	2	2	4	2	2	8
Thr	ACT	5	5	3	5	5	2	13
	ACC	3	11	6	3	11	6	20
	ACA	5	2	5	5	2	5	12
	ACG	2	2	0	2	2	0	4
Ala	GCT	8	13	17	7	12	15	38
	GCC	9	11	8	8	11	8	28
	GCA	8	7	5	8	7	5	20
	GCG	0	2	3	0	2	3	5
Tyr	TAT	3	1	2	3	1	2	6
	TAC	13	5	5	13	5	4	23
His	CAT	10	3	7	10	3	6	20
	CAC	7	6	8	7	6	8	21
Gln	CAA	46	18	28	46	18	28	92
	CAG	23	29	22	23	29	22	74
Asn	AAT	10	12	16	9	11	16	38
	AAC	24	27	15	24	26	15	66
Lys	AAA	10	5	6	10	5	6	21
	AAG	8	2	8	7	2	8	18
Asp	GAT	7	6	13	7	6	13	26
	GAC	9	13	4	9	13	4	26
Glu	GAA	42	26	22	42	26	22	90
	GAG	26	23	25	26	23	25	74
Cys	TGT	7	6	3	6	6	2	16
	TGC	9	3	4	8	2	2	16
Arg	CGT	6	6	4	6	6	4	16
	CGC	8	9	3	8	9	3	20
	CGA	9	6	7	9	6	7	22
	CGG	4	3	5	4	3	5	12
	AGA	12	14	14	12	14	14	40
	AGG	29	16	19	28	16	19	64
Gly	GGT	2	10	11	2	10	11	23
	GGC	6	10	7	5	9	6	23
	GCA	14	10	21	14	10	21	45
	GGG	6	3	7	6	3	7	16

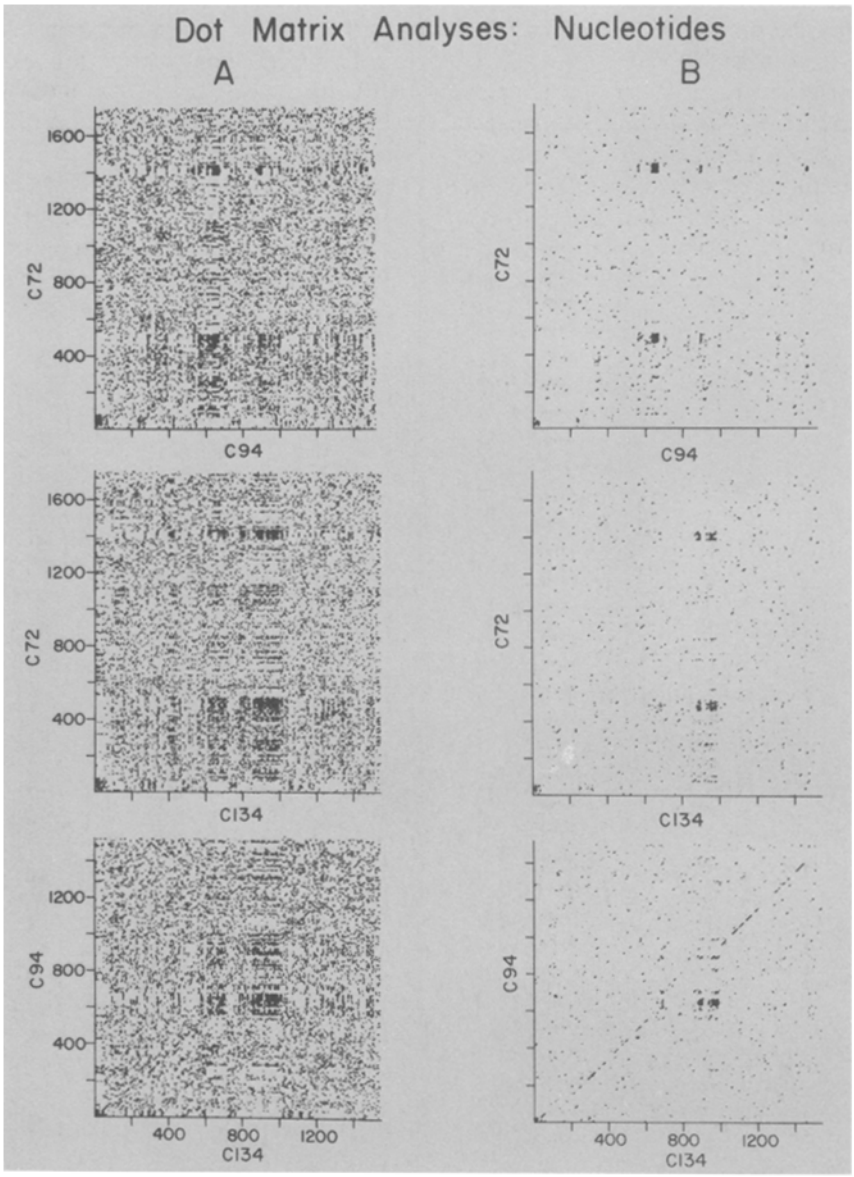

Fig. 6. Comparison of nucleotide sequence of the 3 clones by dot matrix analyses. See text for description of stringencies of comparisons in each panel. Only coding nucleotides are compared.

dot. The increase in span greatly offsets the relaxed stringency and less background is generated, since the random possibility of a 9/12 match is 1 in 2829 sets of 12. Since roughly 1500 sets in one sequence is compared with a similar number in the other sequence, about 800 random dots are anticipated. When C-72 sequence is matched against C-94 or C-134 at either stringency, no regions of homologies are observed, save for the boxes of dots that demark the clusters of Glu found in all these proteins (GAG, GAA clusters). However, when C-94 is compared with C-134 at the low span, a diagonal line from beginning to end of the sequences is discerned. This line of homology becomes more obvious against the low background of dots given by the long span matrix (bottom right panel). Clearly C-94 and C-134 cDNAs are related in their nucleotide sequence throughout and, clearly, sequences in

the middle of the coding regions have moved in time with respect to one another.

Figure 7 gives the same comparisons for the amino acid sequences of the clones. Three comparison stringencies are used here. The first (left panels) compares each amino acid of each sequence and since the random possibility of each comparison is 1 in 20 and since roughly 500 amino acids are being compared, about 12500 random dots are anticipated ($500 \times 500/20$). In the comparisons of C-72 with C-94 and C-134 only the blocks of poly Glu are noticed against the high background. However, the comparison of C-94 with C-134 gives a diagonal line of homology from N termini to C termini of each sequence.

In the middle panels the span of amino acids has been set at 12 and 7 positions must match to generate a dot. Since the random occurence of a 7/12 match is about about 1 in 2×10^6 sets of 12, no random dots are anticipated ($500 \times 500/2000000$).

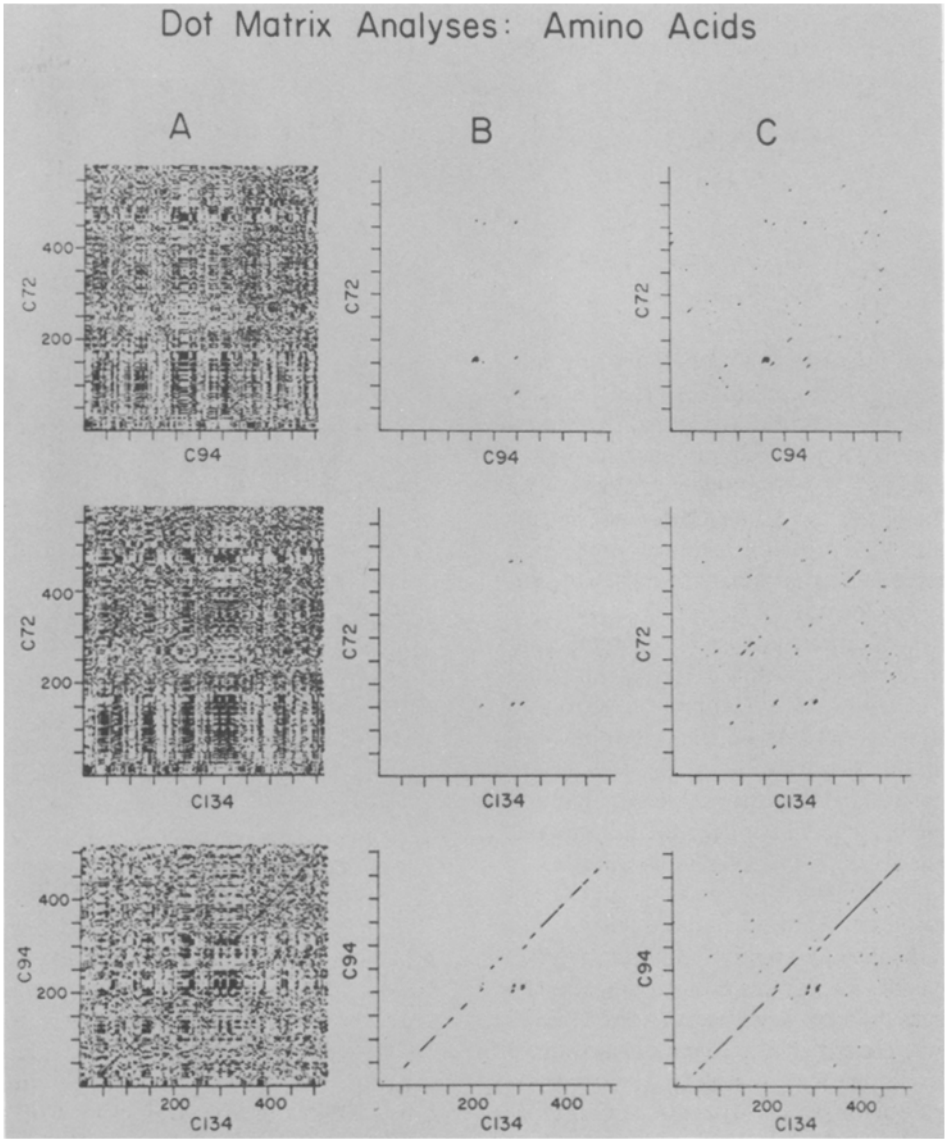

Fig. 7. Comparison of amino acid sequence of the protein represented by the 3 clones by dot matrix analyses. See text for description of the stringencies of comparisons in each panel. The N terminus of each protein is in the bottom left corner of each plot.

This erases all background and leaves only a few dots marking the poly Glu tracts when C-72 is compared with C-94 and C-134. The vestigial homology between C-94 and C-134, however, is clearly shown in this matrix plus the fact that sequences have moved in one of these genes sets relative to the other in the middle of the coding region.

The third comparison again requires that 7 of 12 amino acids match in position but reduces the number of amino acid possibilities to 10 by consolidating what may be considered 'synomyn' amino acids, i.e. grouping as a single entity those amino acids with the R groups that may be functionally similar. These groupings are:

Asp: Glu
Asn: Gln
Arg: Lys
Ser: Thr
Phe: Tyr:Trp
Ala: Val:Ile:Leu:Met

This comparison increases the probability of a random 7/12 matched set to about 1 in 21 000 sets of 12. This predicts about 12 random dots in each matrix. More than 12 dots appear as background in the comparisons of C-72 with C-94 and C-134. This probably is due to the fact some amino acids occur more frequently in proteins than others which would generate a somewhat greater number of random matches, and/or due to the fact that there are preferred tracts of amino acids in seemingly unrelated storage proteins. The effect of this matching of synonyms in the C-94: C-134 comparison is to increase the number of dots on the horizontal homology line indicating that many of the amino acids changes are between functionally synonymous amino acids. This maximizes the demonstration of relatedness of C-94 and C-134 genes.

The vestigial homology between C-94 and C-134 in both nucleotides and amino acid sequence is not obvious when these sequences are compared by eye. Further, there is obviously greater homology on the amino acid level than between nucleotides indicating that many nucleotide changes have not resulted in amino acid changes.

Hydropathy analysis

Gross structural features of the 3 preproproteins are shown by their hydropathy profiles given in Fig. 8. Here the relative hydrophilicity/hydrophobicity of amino acid domains is shown by a summation of the hydrophobicity (5) of spans of 11 (1–11, 2–12, etc.) amino acids (N terminal on the left). Positive values denote hydrophobicity.

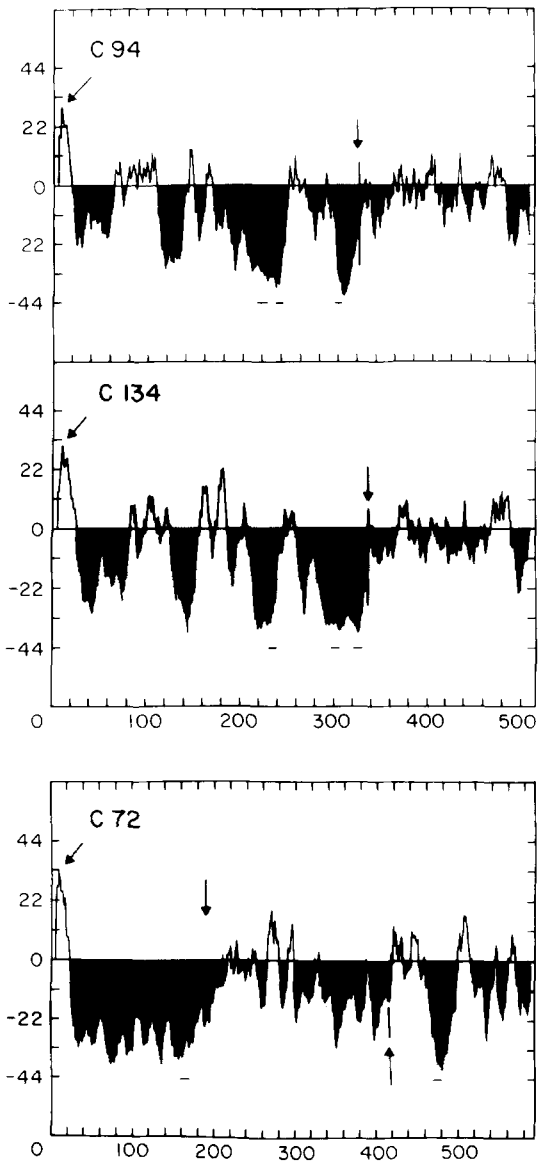

Fig. 8. Hydropathy plots of the amino acid sequences of the proteins represented by the 3 clones. Hydrophilic domains are highlighted in black. N termini on the left. The hydropathic values for amino acid are from ref 5. The span of amino acids is 11. Arrows are explained in the text.

All three preproproteins are extremely hydrophilic. The insolubility of the mature proteins in dilute salts at neutral pH suggest that the proteins are aggregated in protein bodies by salt bridges between charge residues. High salt concentrations (e.g. 0.5 M NaCl) solubilize these proteins (23) presumably by masking these charges.

The N terminal leader sequence is obvious in all 3 instances by their hydrophobicity (shown by left most arrows). The Asn/Gly cleavage sites are shown by downward pointing arrows in C-94 and C-134 as is the putative Arg Arg/Ser cleavage site of C-72, the glycosylation site of C-72 is shown by the upward pointing arrow. Regions of poly Glu are shown by horizontal lines in the hydrophilic portions of the profiles. The profiles of C-94 and C-134 are very similar save in the middle region where a sector containing a clustered Glu tract appears to have moved in one or the other sequence.

Discussion

From the foregoing, it is apparent that the cotton storage proteins represent 2 families of genes – one giving rise to the 69 kD preproproteins and the other to the 60 kD preproproteins (actually 58 kD, see Tables 1 and 2). The latter family is composed of 2 subfamilies that superficially do not seem related. However, sequence dot matrix and hydropathy analyses reveal that the 2 subfamilies emanate from a common ancestral gene. The other gene family that is represented here by cDNA C-72 can also be subdivided into subfamilies A and B based on sequence data from genomic DNA that reveals that subfamily B has a glycosylation site, whereas subfamily A does not. Further, the preproproteins of subfamily B are somewhat shorter than those of subfamily A, having a stop codon 19 triplets upstream from the stop codon of subfamily B (16). The genomic DNA sequences and genome organization data, still uncomplete, will be presented elsewhere.

We suggest that the gene family giving rise to the largest preproproteins be considered the α globulins of the cotton seed and the other gene family be considered the β globulins. The following diagram summarizes the characteristics of these gene families and their A and B subfamilies.

Cottonseed storage proteins

α Globulin family
- 15% of seed mRNA
- cleaved once after leader removed to yield Cys-rich fragment and large mature protein
- has cryptic sequence homology with vicilin-like storage proteins of other angiosperms
- genes arranged as A–B tandems in genome

A Subfamily
- no glycosylation sites
- yields 46.5 kD mature proteins

B Subfamily
- represented by cDNA C-72
- yields 52 kD glycosylated mature proteins

β Globulin family
- cleaved 2× after leader removed to yield 3 mature proteins of 10–25 kD, 2 of which are disulfide linked
- has cryptic sequence homology with legumin-like storage proteins of other angiosperms
- has no glycosylation sites

A Subfamily
- represented by cDNA C-94
- 15% of seed mRNA

B Subfamily
- represented by cDNA C-134
- 5% of seed mRNA.

Some of these characteristics were published in earlier work (1, 2, 11, 23, 24) while still others (disulfide bonding, genome arrangement and sequence relationships with proteins of other plant families) will be published elsewhere (6, 16, 22).

The fact that the precise cleavage points involved in the processing of the proproteins are not known does not prevent the identification of all the small cottonseed storage proteins as products of the cleavages of the 2 families of proproteins. As pointed out in ref 1, characteristics of the mature proteins such as molecular weight, pI, 'negative staining' of poly Glu tracts, kinetics of processing, transient processing intermediates and more recently disulfide linkages all predict the precursor/prod-

uct relationships between the mature species and the preproproteins. Thus, all the cotton seed storage proteins are seen to be derived from the α and β globulins.

The genome of the commercial cotton used in this work is an amphidiploid tetraploid formed by the combination of the A ('old world') and D ('new world') genomes of the cotton tribe (25). An appealing notion would be that perhaps the α globulin family is derived from one of these ancestral genomes and the β globulins from the other ancestral genome. Another possibility would be that the A and B subfamily of each family are derived from the one or the other of the ancestral genomes. However, 1D electrophoresis of the storage proteins from species carrying exclusively the A genome (G. arboreum) and the B genome (G. thurberi) show that existance of both families and both subfamilies of proteins in each genome, albeit with some small differences in molecular weights (data not given). Thus both families and their subfamilies came into existance in the evolution of cotton before the separation of the A and D genomes.

Acknowledgements

This work was supported by funds from the U.S. NSF and Agrigenetics Research Associates Limited.

References

1. Dure LS, Chlan CA: Cottonseed storage proteins: products of three gene families. In: van Vloten-Doting L, Groot GSP, Hall TC (eds) Molecular Form and Function of the Plant Genome. Plenum Press, New York, 1985, pp 67 – 79.
2. Galau GA, Chlan CA, Dure LS: Developmental biochemistry of cottonseed embryogenesis and germination XVI. Analysis of the principal cotton storage protein gene families with cloned cDNA probes. Plant Mol Biol 2:189 – 198, 1983.
3. Maxam AM, Gilbert W: Sequencing end-labelled DNA with base-specific chemical cleavages. In: Grossman L, Moldane K. (eds) Methods in Enzymology, Vol 65, Academic Press, New York, 1980, pp 499 – 560.
4. Garoff H, Ansorge W: Improvements of DNA sequencing gels. Anal Biochem 115:450 – 457, 1981.
5. Kyte J, Doolittle RF: A simple method for displaying the hydropathic character of a protein. J Mol Biol 157:105 – 132, 1982.
6. Borroto K, Dure LS: Sequence homologies among the storage protein of angiosperms. Submitted.
7. Fagan JB, Pastan I, de Crombrugghe B: Sequence rearrangement and duplication of single stranded fibronectin cDNA probably occurring during cDNA synthesis by AMV reverse transcriptase and E. coli DNA polymerase I. Nucl Aci Res 8:3055 – 3064, 1980.
8. Laughon A, Scott MP: Sequence of a Drosophila segmentation gene: protein structure homology with DNA-binding proteins. Nature 310:25 – 31, 1984.
9. Von Heijne G: Patterns of amino acids near signal sequence cleavage sites. Eur J Biochem 133:17 – 21, 1983.
10. Mizuno K, Matsuo H: A novel protease from yeast with specificity towards paired basic residues. Nature 309:558 – 560, 1984.
11. Dure LS, Chlan CA, Galau GA: Cottonseed storage proteins as a tool for developmental biology. In: Ciferri O, Dure LS (eds) Structure and Function of Plant Genomes. Plenum Press, New York, 1983, pp 113 – 121.
12. Chilton WS, Tempé J, Matzke M, Chilton M-D: Succinamopine: a new crown gall opine. J Bacteriol 157:357 – 362, 1984.
13. Ullrick A, Coussens L, Hayflick JS, Dull TJ, Gray A, Tam AW, Lee J, Yarden Y, Liberman TA, Schlessinger J, Downward J, Mayes ELV, Whittle N, Waterfield MD, Seeburg, PH: Human epidermal growth factor receptor cDNA sequence and aberrant expression of the amplified gene in A431 epidermoid carcinoma cells. Nature 309:418 – 425, 1984.
14. Ebina Y, Ellis L, Jarnagin K, Edesy M, Grof L, Clauser E, Ou J, Masiarz F, Kan YW, Goldfine ID, Roth RA, Rutter WJ: The human insulin receptor cDNA: the structural basis for hormone-activated transmembrane signalling. Cell 40:747 – 758, 1985.
15. Hollenberg SM, Weinberger C, Ong ES, Cerelli G, Oro A, Lebo R, Thompson EB, Rosenfeld MG, Evans RM: Primary structure and expression of a functional human glucocorticoid receptor cDNA. Nature 318:635 – 641, 1985.
16. Chlan CA, Borroto K, Kamalay J, Dure LS: Genome organization of the storage proteins of cottonseed. Submitted.
17. Moreira MA, Hermondson MA, Larkins BA, Nielsen NC: Partial characterization of the acidic and basic polypeptides of glycinin. J Biol Chem 254:9921 – 9926, 1979.
18. Casey R, March JF, Sanger E: N-terminal amino acid sequence of Beta-subunits of legumin from *Pisum sativum*. Phytochem 20:161 – 163, 1981.
19. Lycett GW, Croy RRD, Shirsat AH, Boulter D: The complete nucleotide sequence of a legumin gene from pea (*Pisum sativum*). Nucl Acid Res 12:4493 – 4506, 1984.
20. Simon AE, Tenbarger KM, Scofield SR, Finkelstein RR, Crouch ML: Nucleotide sequence of a cDNA clone of *Brassica napus* 12S storage protein shows homology with legumin from *Pisum sativum*. Plant Mol Biol 5:191 – 201, 1985.
21. Walburg G, Larkins BA: Oat seed globulin. Subunit characterization and demonstration of its synthesis as a precursor. Plant Physiol 72:161 – 165, 1983.

22. Borroto K, Dure LS: Disulfide linkages between the β globulin storage proteins of cottonseed. Submitted.
23. Dure LS, Chlan CA: Developmental biochemistry of cottonseed embryogenesis and germination XII purification and properties of the principal storage proteins. Plant Physiol 68:180 – 186, 1981.
24. Dure LS, Galau GA: Ibid XIII Regulation of the biosynthesis of the principal storage proteins. Plant Physiol 68:187 – 194, 1981.
25. Fryxell PA: The Natural History of the Cotton Tribe. Texas A & M University Press, College Station.

Received 12 May 1986; in revised form 31 July 1986; accepted 6 August 1986.