Characteristics of the Storage Proteins of Cotton

Leon Dure III
University of Georgia, Department of Biochemistry, Athens, GA 30602

The storage proteins of the cottonseed are composed of α and β globulins which are homologous to the vicilins and legumins, respectively, of the legumes. Gene sequences for members of both protein families have been determined, and the derived amino acid compositions are presented here. The proteolytic cleavages and glycosylations involved in the processing of the preproproteins during embryogenesis to yield the proteins found in the mature seed are outlined.

The storage proteins of the cotton seed (*Gossypium hirsutum*) are composed of two protein families, the α and β globulins. At seed maturity these proteins constitute about 60% of the embryo protein and are contained exclusively in protein bodies. They emanate from two families of genes, undergo several endoproteolytic cleavages and, in the case of one subfamily, are glycosylated at a single site as they move from preproproteins to the mature proteins of the seed. This processing gives rise to the storage protein profile on SDS-PAGE shown in Figure 1, lane 3. In this lane, which is of the total protein extracted from a partially purified preparation of protein bodies, eight distinct protein bands can be discerned. The largest proteins (~51 and 48 kD) are α globulins, whereas the smaller proteins, labelled 1–6, are principally products of the β globulin gene family. Only the 51 kD proteins are glycosylated. The faintly stained proteins in this lane are not storage proteins but contaminants of the protein body preparation. These bands give an over-simplified view of the mature proteins when compared to a 2 D presentation of this same preparation (Fig. 2). It is apparent that each band in Figure 1 is actually composed of multiple polypeptides that differ in isoelectric point. Thus, these proteins emanate from multigene families.

When mRNA, isolated from developing cotyledons, is translated in the wheat germ system and the products displayed on SDS-PAGE, two major bands of large unprocessed proteins are seen (Fig. 1, lane 1). These two groups of proteins are the preproproteins of the α (70 kD) and β (58 kD) globulins, and they give rise to all the proteins seen in lane 3 of this figure. Some of the intermediates formed as the preproproteins are processed to the mature proteins accumulate during embryogenesis and can be seen on stained protein gels (Fig. 1, lane 2). Here, protein bands of 71, 68 (faint) and 35 kD are seen that are not present in the mature seed. The 71 kD species is glycosylated.

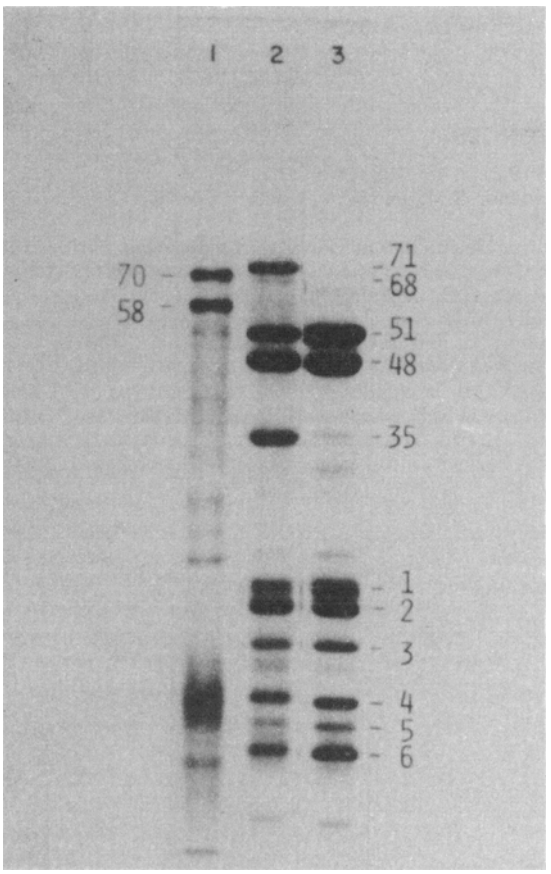

FIG. 1. SDS-PAGE of cotton seed proteins. In lane 1 wheat germ translation products of mRNA taken from mid-maturation stage cottonseed cotyledons have been fluorographed and aligned with stained proteins obtained from partially purified protein bodies taken from cotton cotyledons of mid-maturation stage (lane 2) and of the mature seed (lane 3). The apparent molecular weights of the large proteins are given in kD; the bands containing smaller storage proteins are designated 1–6.

PROCESSING EVENTS IN STORAGE PROTEIN MATURATION

We have isolated, cloned and sequenced several genes from the two storage protein families (1,2) and have

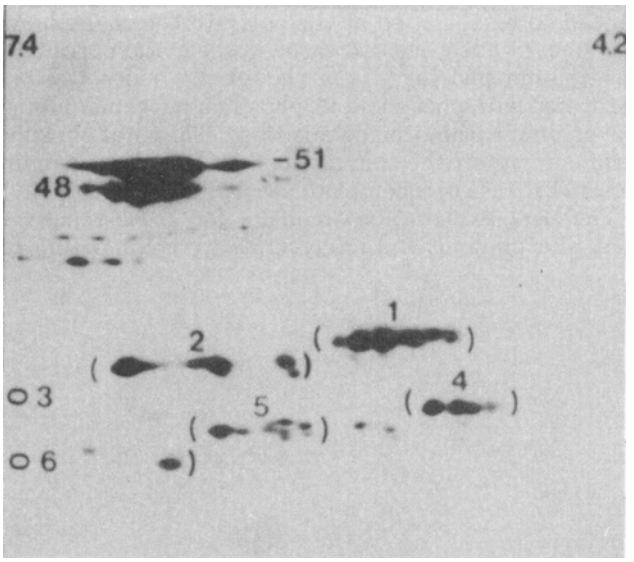

FIG. 2. A stained two-dimensional gel electrophorogram of protein body protein obtained from cotyledons of mature cottonseed. The pH range of the isoelectric dimension is given at the top of the gel; the small protein bands of Fig. 1 are identified as 1–6. The proteins of bands 3 and 6 have migrated off the alkaline end of the gel due to their basic pI.

been able to construct the processing pathways that result in the array of mature proteins found in lane 3 of Figure 1. Data from a number of different experiments has allowed us to accomplish this.

Because these proteins are all contained within protein bodies, a leader sequence must exist in the preprotein translation products to allow for passage through the ER en route to protein body disposition. The point of cleavage of the leader sequences is readily determined by the rules of von Heijnes (3).

We noted from earlier studies the immunologic cross reactivity between the 51 and 48 kD mature proteins (4).

From gene and cDNA sequences we have established through sequence homology that the α globulin family is a member of the "vicilin" family of globulin storage proteins that is common throughout the dicots. Likewise, the β globulin family of cotton is a member of the ubiquitous "legumin" family (5). There is no conserved pattern of processing of vicilin proteins among dicots, but there is a common cleavage site among legumins that is found in the cotton β globulin sequences and was useful in predicting precursor-product relationships.

Of the storage proteins only the 71 kD processing intermediates and the 51 kD mature proteins are glycosylated, which clearly shows the 71 kD species as precursor to the 51 kD species. Sequence analysis of the α globulin genes shows that some of the α globulins have a single glycosylation site (asn val thr), whereas other α globulins do not. Hence, the nonglycosylated α globulins of 48 kD result from the processing of the 68 kD intermediates.

Because some of the processing intermediates accumulate in embryos, their final cleavage products should accumulate more slowly than others. Thus, pulse labelling studies following the kinetics of accumulation of mature species allowed us to equate specific mature products with these intermediates (6,7).

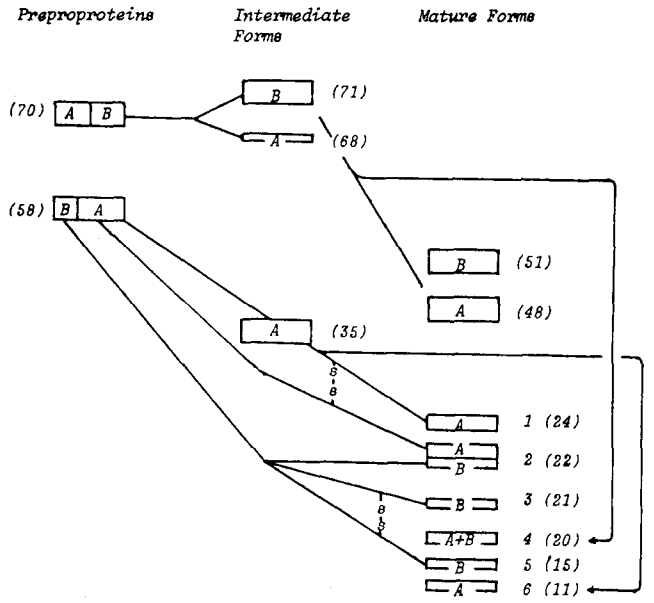

FIG. 3. A diagram showing the processing events that convert the α and β globulin preproproteins to the mature species of the dry cottonseed. The small protein bands listed as 1–6 equate them with Fig. 1. The 71 and 51 kD species are glycosylated, and the mature proteins that are disulfide-linked are shown.

Usually the cleavage products of legumin processing are covalently linked to each other by disulfide bonds even after the polypeptide chain has been cleaved. By running gels ± reducing agents we were able to demonstrate that certain of the mature products of the legumin family were so linked (8).

All of these observations have allowed us to follow the formation of the mature proteins from their preproprotein progenitors as diagrammed in Figure 3. This figure shows the α globulin family to be composed of two subfamilies, an unglycosylated A subfamily and a glycosylated B subfamily. The preproproteins of both these subfamilies are about 70 kD. After losing the leader sequence the A subfamily proproteins of about 68 kD are only faintly visible on stained genes (lane 2) due to their rapid cleavage to give the 48 kD mature species. The leaderless B subfamily, on the other hand, becomes glycosylated and migrates as species of about 71 kD. Their slow cleavage to yield the 51 kD species allows them to become quite abundant during embryogenesis. The other fragments produced from the cleavage of the proproteins of this family of proteins are about 20 kD that comprise band #4. The precise point of this cleavage is unknown, but indirect evidence places it between amino acids 180–200 from the N terminal of the preproproteins (1,5).

The β globulin family is also composed of two subfamilies that, although clearly members of the legumin group of storage proteins from sequence analysis (5), have diverged considerably from each other. The β globulin A subfamily proproteins, after losing their leader sequence, are cleaved to produce fragments of about 35 and 22 kD that are linked by a disulfide bond. The 35 kD fragment is slowly cleaved a second time to yield a 24 kD mature protein, that is still linked by a disulfide to the 22 kD piece, and a small piece of about 11 kD. The sizes of the 35 kD intermediate and the 22 kD protein are known precisely from sequence analysis. However, precise sizes of the 24 and 11 kD pieces are surmised from gel electrophoresis because the site of cleavage is unknown. In the diagram the mature proteins of band 1, part of band 2 and band 6 are shown to emanate from the A subfamily preproprotein.

The B subfamily proproteins give rise to a 21 kD mature protein whose precise size is known, and a larger piece that must be rapidly cleaved again to yield mature proteins of about 22 and 15 kD. Again, these sizes are estimates because the site of cleavage is unknown. The 21 and 15 kD species are disulfide linked and are bands 3 and 5 on the 1 D gel. The pI of the 21 kD protein is too basic to be retained on the 2 D gel. The 22 kD proteins comprise a portion of band 2.

The initial cleavage of the proproteins of the β globulins is a property of all angiosperm legumin storage proteins, occurring at the highly conserved site mentioned before. The second cleavage of the large fragment is unique to the cotton legumins at this point.

AMINO ACID COMPOSITION

The amino acid composition of the proproteins of each of the subfamilies of the two groups of proteins is

TABLE 1

Representative Amino Acid Composition of the Cotton Storage Protein Families

	α globulin family (vicilins) (proproteins)		β globulin family (legumins) (proproteins)	
	A gene subfamily	B gene subfamily[a]	A gene subfamily	B gene subfamily
ASP	25	17	19	17
ASN	34	31	37	31
GLU	65	72	49	47
GLN	63	67	47	50
ARG	63	62	54	52
LYS	24	20	7	14
HIS	15	16	9	14
SER	34	38	30	30
THR	18	16	20	13
CYS	14	14	8	4
PRO	27	30	20	22
GLY	37	29	32	45
ALA	23	24	32	31
VAL	37	27	27	30
ILE	12	18	23	21
LEU	32	29	31	32
PHE	33	34	22	22
TYR	17	15	6	6
TRP	5	3	6	7
MET	2	3	9	6
TOTAL				
residues	580	565	488	494
Mol wt	68,280	66,863	56,148	56,352

[a]This subfamily is glycosylated and gives higher apparent molecular weights on electrophoresis.

given in Table 1. Because the cleavage sites for producing many of the mature proteins have not been precisely determined, an exact amino acid composition is not possible for these mature forms individually, although in summation they should total the composition of their proproteins.

PERIOD OF SYNTHESIS

It takes about 50 days from anthesis for this cotton cultivar "Coker 201" to reach boil opening and the commencement of desiccation. In the first 20 days of this period the embryo remains small but develops a recognizable axis and cotyledons. Endosperm and nucellus growth are extensive during this period. Between days 20 and 50 the embryo grows rapidly at the expense of the endosperm/nucellus, going from about one mg to about 125 mg in wet weight. By day 50 the endosperm and nucellus have been consumed and the seed begins to sclerify, and between days 50 and 55 the embryo desiccates to the mature seed with a wet weight of about 65 mg.

We have detected a low level storage protein mRNA and storage protein accumulation in the young one-mg embryos. Messenger RNA and storage protein synthesis increase to reach a maximum level at about the 40-mg stage. This level is maintained until a few days before desiccation begins (about 110 mg stage), at which time the storage protein mRNA drops precipitously over a 2-3 day period. Only traces of these mRNAs exist in the mature seed. During the period of maximum synthesis, storage protein mRNA appears to represent about 35% of total cotyledon mRNA (9). This seems to be distributed as 15% α globulin mRNA, 15% β globulin A subfamily mRNA and 5% β globulin B subfamily mRNA. These values must be considered approximations only because they were determined by solution hybridizations of cDNA and mRNA, which are not overly accurate. In a general sense it would seem from stained protein gels that the α and β globulins contribute equally to the total storage protein of the seed. The two subfamilies of the α globulins appear to contribute equally to the α globulin fraction. However, as for the β globulins, there appears to be much less of the B subfamily proteins than of the A subfamily on stained gels. As mentioned above, there is less B subfamily mRNA than A subfamily mRNA during the period of maximum synthesis. It is not known if this reflects the existence of more A than B subfamily genes in the *G. hirsutum* genome. Based on RNA dot blots the expression of the α and β globulins seems to be coordinate during embryogenesis (8).

GENETICS AND GENOMIC ORGANIZATION

G. hirsutum is a naturally occurring allotetraploid containing both the "old world" A genome and the "new world" D genome (9). Gel electrophoresis of species containing only the A or the D genomes shows that both genomes contain both subfamilies of the α and β globulins. The size of the mature proteins differs somewhat from that of *G. hirsutum*, indicating that the processing cleavage sites have moved in the genes as these species have evolved. Thus the A and B subfamilies of both globulin families diverged prior to the divergence of the A and D genomes.

We have found that the α globulin genes exist in the *G. hirsutum* genome as tandems of B gene — spacer — A gene (as oriented in the direction of transcription). Three different tandems have been mapped with restriction endonucleases, and all are different from each other and from the map of a cDNA for a B gene that has been sequenced, which suggests that there are at least four such tandems. This would allow for two

tandems to have been supplied by the A genome and 2 by the D genome. Based on sequence analysis, the B genes, which contain the glycosylation site, differ among themselves only about 2% in their nucleotide composition and 4% in their amino acid composition. An A gene differs considerably more from a B gene, showing an 18% nucleotide divergence in the coding regions and a 28% divergence in amino acid sequence (2).

As yet we know nothing about the number of genes, nor their organization, for the β globulins other than that no fragment of genomic DNA carrying an A or B subfamily gene has another storage protein gene within eight kb up or downstream from it.

REFERENCES

1. Chlan, C.A., J.P. Pyle, A.B. Legocki and L. Dure III, *Plant Mol. Biol. 7*:475 (1986).
2. Chlan, C.A., K. Borroto, J. Kamalay and L. Dure III, *Ibid. 9*:533 (1987).
3. Von Heijne, G., *Eur. J. Biochem. 133*:17 (1983).
4. Dure, L. III, and C.A. Chlan, *Plant Physiol. 68*:180 (1981).
5. Borroto, K., and L. Dure III, *Plant Mol. Biol. 8*:113 (1987).
6. Dure, L. III, and G.A. Galau, *Plant Physiol. 68*:187 (1981).
7. Dure, L. III, and C.A. Chlan, in *Molecular Form and Function of Plant Genomes*, edited by L. van Vloten-Doting, G.S.P. Groot and T.C. Hall, Plenum Publishing Corp., 1985, pp. 67-79.
8. Dure, L. III, J.B. Pyle, C.A. Chlan, J.C. Baker and G.A. Galau, *Plant Mol. Biol. 2*:199 (1983).
9. Fryxell, P.A., *The Natural History of the Cotton Tribe*. Texas A&M University Press, College Station, TX.

[Received August 11, 1987; accepted October 7, 1988]