

Plant Gene Register

Cotton *Mat5-A* (C164) Gene and *Mat5-D* cDNAs Encoding Methionine-Rich 2S Albumin Storage Proteins¹

Glenn A. Galau*, Helen Y.-C. Wang, and D. Wayne Hughes

Department of Botany, University of Georgia, Athens, Georgia 30602

Several abundant mRNAs are coordinately expressed specifically during the maturation stage of embryogenesis (6, 7). The cotton *Mat* mRNAs include those of the major vicilin and legumin storage protein genes (6), and representatives of these genes have been sequenced (2, 3). To help define the temporal and spatial regulation of the maturation program of gene expression, we isolated an additional cotton *Mat* gene, *Mat5*, which encodes mRNAs represented by cDNA clone C164. The mRNAs from the two *Mat5* alleles in the allotetraploid genome of *Gossypium hirsutum* make up 2% of the mRNA in maturation stage embryos (6), and the sequences show that they encode methionine-rich 2S albumin storage proteins.

Three genomic fragments were recovered, but restriction site analysis and sequencing showed that they contain the same allele, *Mat5-A*, present in the A genome. One was extensively sequenced, as was all of cDNA C164 and extensive portions of five other cDNA clones (Table I). Figure 1 shows the sequence of clone GC164–24RC of *Mat5-A*. All six cDNAs are identical with each other in the regions sequenced, but they differ significantly from the sequence of *Mat5-A*. The gene and cDNA sequences are colinear, with eight nucleotide differences, until *Mat5-A* nucleotide 3661 and C164 nucleotide 514. There is no obvious similarity for the next 62 nucleotides until the site of polyadenylation of the cDNAs (Fig. 1). Because there are only two allelic genes in *G. hirsutum* (6), we presume that all the sequenced cDNAs are transcribed from the other allele, *Mat5-D*.

The two *Mat5* alleles encode proteins that differ in only three of 139 amino acid residues. Their sequences are very similar to those of the 2S albumin storage protein family (1, 8, 9), and an alignment of the cotton sequences with other 2S albumin preproteins is unambiguous in its important features (Fig. 2). There is complete identity at the cysteine and leucine residues that are diagnostic of the family and about 44% of the cotton residues are identical with those of *Bertholletia* (1) and *Arabidopsis* (9) 2S albumins. The predicted mature form (9; Fig. 2) of the cotton 2S albumin shares with

that of *Bertholletia* a high methionine content. It is 10% in cotton and 19% in *Bertholletia*, compared with 2 to 3% in other 2S albumins (1, 8, 9). Four of the 10 methionine residues in the cotton protein are at novel positions that may be sites at which other 2S albumins could be engineered for higher methionine content.

Because none of the six cDNAs that were examined are transcribed from *Mat5-A*, it is possible that this gene is not active. However, it is more likely that the apparent absence of *Mat5-A* cDNAs is due to failure to process or polyadenylate its transcripts. In the sequence in Figure 1 and in overlapping phage isolate GC164–27 (Table I), there is a class I-like retrotransposon (10) in reverse orientation to the *Mat5-A* transcription unit (data not shown). The retrotransposon's putative internal domain/3'-long terminal repeat boundary is at *Mat5-A* nucleotide 3791/92, and the 130 nucleotides between this boundary and the *Mat5-A/Mat5-D* cDNA divergence point can contain only a highly truncated long terminal repeat, if any part of it. We predict that the element, or element-induced rearrangements of *Mat5-A* sequence, interferes with the processing of the 3' end of *Mat5-A* transcripts. Rearrangement of *Mat5-A* in this region is suggested by the presence of long direct repeats, part of one of which is present in *Mat5-D* cDNAs (Fig. 1).

A sequence very similar to *Mat5-A* nucleotides 907 to 1765 has been found in reverse orientation in the upstream region of an unrelated cotton *Lea4* gene (4). Portions of this element are also duplicated at *Mat5-A* nucleotides 2289 to 2416 (Fig. 1, repeats 1 and 2). Consequently, the sequences that are responsible for transcription of *Mat5-A* are probably between nucleotide 2615 and the transcription start at nucleotide 3146. In this region the sequence of *Mat5-A* has some similarities with those of other 2S albumin storage proteins, in particular two elements (double underlined in Fig. 1) present in similar locations in many of these genes (8, 9). Other elements suggested to be important in various storage protein genes (5) are not obvious in *Mat5-A*. Finally, *Mat5-A* has been compared with two other cotton *Mat* genes, vicilin (2) and legumin A (3), with which *Mat5-A* is coordinately expressed in cotton (6, 7), but no similarities are compelling in any sequence pair.

¹ Supported by a grant from the National Institutes of Health.

	20	40	60	80	100	
	ATCGATTAATTTTTAATG	TTTAATTAATAATTAATTA	TTCTCAATTAAGTATTTAT	CACCCCTAGCTAAGAATATC	CTTCTGTCTCATCTTCACT	0100
	AAGGATTACCACCTAAGTCA	CTATTCCGGTCTCTCCTAG	GTATACGAATACTCTTTTGA	TCTCACTACTTGGCACAAGT	TATACACAACAAGATACCC	0200
	TCGGGTGCACCTATCTAGA	TATTCTGCCAGGGGTGTAC	TCCGTTATATACTAAGTACT	CTTGAATAAACAACCAATTT	CAAAATAAATAACTACTTAAC	0300
	AAATAAAATGAATCATGCA	AGATGTAATAAAGCACAAT	AATCTATCTACTATTTATA	TAAACAGTAATTGACTAGAT	TAACGAAAAGTAAATAAAAG	0400
	AATAGAAATAGAAGAGATAA	AGAATGCTCATGAATAAATA	TATTGAAGCGTGATCCGGA	GAAATCTCAGCTTCAAATG	TCTTCACATCGAAGAGAGGT	0500
	TCCTACGAAGAACATCAAG	AAAAAATACAAAGAAATTTG	ATTAGTAAAAATGTCTGTCT	TAAGTTTGGCCACCATTGAG	GTGATCCATAAGTACTTAT	0600
	ATATGCTACGGGTGACTTGG	TGACAATTTAACCTAGACAC	ACTTTGGGATTTCTAATTAAT	GCAATATTTTGATAAATGCT	TGGAGTCCAAATAATGAAAT	0700
	ATCTCAAAAGTGTCATCCAA	GGAATCAACTTCATTTTAAT	TTGCTTCATTCGGCTCATCG	TGACGTTGGGAACTCCTTGT	CACGTCACAATGTCGAATGA	0800
	ATTTTTAGCCCTTTGTCAATTT	TGTTGTTCCATGTTGCAACA	CGAGCTTCTTGTATCTCGA	CCCGCAGATTGTTTCTTGA	TTTGTGTAGTATAACTCATA	0900
	TTTCAGTCAAAATTAAGAGA	TAATCTCCCAGTTAACTTTT	GTTTATTTGTTTAATCGAA	CTCTGATTAGTCTAGATTAT	TTTATTTGTTTGGACCTAC	1000
	GAATTTAGCTATAAATTTTG	TTCTTTTACAACCTTAGAAA	ATACACCCATTAAGATCTG	AGCTCATAACTCTTTAGAG	AATTTGTGTTACTGTTTGG	1100
	AAGATTTTTGTTCGAGT	TTCCGGGTTTGTTCGAGT	CTCCATCTTTTACTCTTC	GTTCTTTTGTCTATTAGTA	AAATTTATCTTTACTCGTAGA	1200
	<u>TTTTATCCTCTTGGAGGG</u>	<u>GTTTTCCACGTTAAATTTG</u>	<u>TGTATTTTTTCGCTATGTT</u>	<u>TACTTGTTCGTTGCTTAATC</u>	<u>AGGTCGATCCCAACAAGTG</u>	1300
	GTATCAGAGTTAGTTCAAT	TTCATAGATCAACCCATTTA	GAAATGTCAACAATAAGGTT	TGACATTGAGAAGTTCGATG	GTGTCACAAATTTCAATATG	1400
	TGGTAAGTTCGAATATGAC	AATCTAGTCAAACTCGGCC	TGAAAAGGTTGTACTTGG	AAAAAGCTTGAATACTAGA	TAAGACAATAATGGCAGAGT	1500
	TGATGAAAAGGCTCTATCTG	CAATTCATTTGTCTCTCGG	AATATGGTATTCGAAAAGTT	ATTGATGGAGAAAACCTCAT	TCGCCTTGTGAAAAGATTA	1600
	AAAATCTTTATGCGAATAA	GTTCTAGCTAACCTTTTAG	TATTGAAATGACGCTTATTT	ACGTTTCGCATGAACGAAT	TGAGCTCTTAACATTCACA	1700
	TTAGTCAATTTATCTCTTT	TTGAATGATTAAGAATAT	TGAGGTTAAGATGACGATG	AAGATAGACCTGTTTCATGGG	TTGGCCACCAGCCTGCTC	1800
	CAAAATGTGGGAGGATTTTG	TAAAAATATAGGCTCGAAA	ATGAGTTTAAACAATAAATG	TAAAAAAAAGGGTCTTATC	TCAGGTAAGGCTTTTGGC	1900
	TCAGGTTGCTACTGTTTTC	ATTCACTAAAGGAAAAAAA	ATGTTATTTTTTAACTACT	ATTTTCTTGTGTTTCTCC	CTATTTGCTACTATTTTAT	2000
	TATTATGTTGCTACTATTTT	GTTGTTATTGTTGGATATT	GTATAAACTTATTTTATG	ATAATTTTGTATTATTTTA	GAGGCAATTTTAAATTTTG	2100
	<u>TATTATTTAGGGCATTTC</u>	<u>CTTGCTAAGTTCATCTATC</u>	<u>TTAGTGTATTTAAGTATAC</u>	<u>ATATTTTTTAAATTTTATTT</u>	<u>TCAATTTGTTGGAAAATAT</u>	2200
	<u>TTTTGATATTTTATATAT</u>	<u>TTTTGATGAATTTATATAT</u>	<u>AAAAAATATAAAAAATTA</u>	<u>TACAGACAAGTCGGATCGGG</u>	<u>CCCGGTTTTTACTTTGCCA</u>	2300
	<u>GTGATTTTTATCTCTTTG</u>	<u>GAGGGGTTTTCCCATGTTAA</u>	<u>ATTTGTGTGTTCAATTTCTC</u>	<u>AATTTCTTCTCTATTTTTA</u>	<u>CTTATTGTTGCTTAATTTGA</u>	2400
	<u>GTGATCCCCAACAGATTC</u>	<u>TTGACCTGATTTGACATCCT</u>	<u>ACACACTTAATTAGCCATTA</u>	<u>ATCAGTATCAAGGCCTAAAT</u>	<u>TGACCTTTACAGGTCATTGA</u>	2500
	<u>GACATCAAAAAATACGTA</u>	<u>AATATTTATTTTATTAATA</u>	<u>TTAAATCCATAAATTA</u>	<u>ACTGAAAAACAATAAATA</u>	<u>TTATAAAAATCCATAAAAA</u>	2600
	<u>AAACTCATTAGGTTAAAA</u>	<u>TGATCTAACTACCCTACT</u>	<u>AGTTGGTGGAAAATAA</u>	<u>GACCCACTTTTTAAGTCATA</u>	<u>AAAGGTGGTTGTATTTTC</u>	2700
	<u>ATAGCCATGCAATGTTAGG</u>	<u>GAAGTGAGATGTTGGTGAG</u>	<u>TCATCTTTGAAAGTTGACCTA</u>	<u>ATGAAGTCTGTGAGAGTAAC</u>	<u>GATCAACTTAAGGAGGTTT</u>	2800
	<u>CCAAGCACCGCAAACCTTT</u>	<u>TTATGATATTTTTTAAAT</u>	<u>TTTGAACATATACATGGGCT</u>	<u>TGAAAAATATCCAATTTG</u>	<u>TTAAGTGTGTTGGGTTGAT</u>	2900
	<u>TCCAATTTGGATCACCGAT</u>	<u>CAAAGTGAACCATTTTATC</u>	<u>CCCAGGAGATTTAAATA</u>	<u>CATCGAGAACCCTGAATACT</u>	<u>CCGATCCCAAGTTAAGGCAGT</u>	3000
	<u>TTCAATAACCAACAAACAA</u>	<u>TGTGTCCATTTTGCATGCA</u>	<u>GAAATTAACCTACGTTAG</u>	<u>GTTTCAAGTTTCAACATCAC</u>	<u>ACGTTATCCCATGCAAAAG</u>	3100
cDNA		transcription start ↓	start cDNA	C103, C164	C	053
gene	TTCAACTCTCTATAAATTA	CCTTCACACCTCCACTCCAT	TTCTCACCCCTTCTCATCTGA	TTATTTTCTCCATACCAGGA	TAAATCAACAATGGCAAAGC	3200
gene					MetAlaLysL	003
cDNA		G			C	153
gene	TCGCAGTTTACTTAGCCACC	CTTGCTCTCATTGTTCTCT	TGCCAATGCTTCAATCACAT	CTGTGACGGTCGAAAGCGGAG	GAAAAATCGGGATAGCTGTGA	3300
gene	euAlaValTyrLeuAlaThr	LeuAlaLeuIleLeuPheLe	uAlaAsnAlaSerIleThrS	erValThrValGluSerGlu	GluAsnArgAspSerCysGl	036
cDNA				Tr		253
gene	ACAGCAGATAAGGAAGCAAG	CCCACCTGAAGCACTGCCAG	AAGTACATGGAGGAAGAGTT	GGGTGGCGAAGGCAGCGACA	ATATAGCCGGCGGGTACATT	3400
gene	uGlnGlnIleArgLysGlnA	laHisLeuLysHisCysGln	LysTyrMetGluGluGluLe	uGlyGlyGluGlySerAspA	snIleAlaGlyGlyTyrIle	070
cDNA					Glu	353
gene	GACTCATGTTGCCAGCAGCT	AGAGAAGATGGATACGCAAT	GCAGATGTCAAGGTCTAAGA	CATGCAACGATGCAACAGAT	GCAACAGATGCAAGGACAAA	3500
gene	AspSerCysCysGlnGlnLe	uGluLysMetAspThrGlnC	ysArgCysGlnGlyLeuArg	HisAlaThrMetGlnGlnMe	tGlnGlnMetGlnGlyGlnM	103
cDNA			Asn			453
gene	TGGGGAGCAAACAGATCGCA	GAGATCATGCAAAAAGGTTAC	CAAGAAAATAATGTCCGAAT	GTGAGATGGAGCCTGGGAGG	TGTGACACGCCATCTCGCAG	3600
gene	etGlySerLysGlnMetArg	GluIleMetGlnLysValTh	rLysLysIleMetSerGluC	ysGluMetGluProGlyArg	CysAspThrProSerArgSe	136
cDNA		A				553
gene	TTTGATTTAGACAAAATATG	TTCTCAAGAATAAATGTAAT	GTCTCCTTTAGTTACTGTGG	GTTTTAGCTGTACTCGCTC	CTTTCATCAAAAAGGCAAGC	3700
gene	rLeulle			<u>TAAACTTGACCACCTTTTCA</u>	<u>AAGTCTTCTTAGGGTGCTA</u>	139
cDNA		end cDNA C164	C87, C90, C103			575
gene	ACATAAGTGTGCAGTATTG	TT poly(A) cDNA C42,	CCCATTAATGTTAAACAATCT	CCACCTGAAGATTTGATTA	GGATAATCACATCTTCACAC	3800
	<u>AACTTGCACCACTTTGTTG</u>	<u>GTTTGCATTAATGTTGCT</u>	<u>CCCATTAATGTTAAACAATCT</u>			
	ACTTCTTCAACTCTCCAAA	TTCGATAAAGTATCTTTTG	TAGTGCTTCAAAATGCACCT	TCGAGCGCCATACACCTGAA	GGTGCTCAAATTTCTCAGGAT	3900
	GTAAATCAAGTTCAAACAA	GATTAACACTGATTTGTTG	ACCACCTTGGTCATCATATC	TCGGGATTATCTGCTGTGC	GAATCTTCTGAAGTAGAATT	4000
	TTTCTTTTTCAAAGACTT	CCGCACAAAGTGATATCTTA	CGTCGATATGCTTGGTCTT	GAATGATAGACTTGATTTT	CGCTAAATGAACAGCGCTCT	4100
	GACTGTCAACAATAAACTT	ATGTGACTTTGAACAACTCC	TAAGTCTTTCAACAATCCAT	TAAGCCAAATAGCCTCCTTA	ACAGCTTGTAACTGCCAT	4200
	ATATTTCTGCCTCTGTAGTAG	ACACAGCTACTGTAGACTGT	AAGGTAGACTTCCAACCTCAC	TGGGGCTTTCCGAAGAGTAA	ACAGATACCCCGTAGTTGAA	4300
	CGACGTTTATCTAAATCAC	AGCAAAGTCGGAATCAACAT	ATCCAACTACAAACTGACCA	AGTGCTTCACTCTGTTCAAA	AATTAACCAACATCTACGG	4400
	TTTTTCGAAGATACCGTAGA	ATCCATTTTACAGCTTGCCA	ATGTCCTTTTCCAGGATCAT	GCATATACCTGCTCACAAC	CCAACAGTGTGAAATGTC	4500
	AGGCCTCGTACACACCATCG	CATATATCAAACTCCCAACT	GCATTAGCATATGGGACTTT	CCCATATATTTCTTTTCT	CTTCAGTTTCCGGAGATAA	4600
	TGAGCACTAAGTTTCAAATG	AGAAGCAAGTGGGACTTAA	CATGTTTGTGTTTTCATTT	ACACCAAAACATTGTAATAC	CTTTTTCAGATATTGCTTCT	4700
	GATTTAAACAGAGCTTGCTT	CTCGGTCTATCTTACTTAT	CTCCATGCCGAGAATCTTCT	TGGCTCACCTAGATCTTTC	ATCTCGAAGTCTTAATTCAA	4800
	CTGAGCCTTCAGCTTATCTA	TGTCATTTTGGCTCTTCGAA	GCGATTAAACATATCATCAAC	ATACAAGAGTAGATAAATGA	AAGATCCCGTCATCGAGCTTC	4900
	TGCAAAATACACAATTTGTC	ATATTTGCTTCTGTGTAAT	CTCGCTTCTCATAAAGCTG	TCAAATCGCTTGTACCAGT	CCTCGGGGATTTGCTTCAAT	5000
	CATATAGCGATTTGTTCCAG	TTACAACCCCAATTTCTACC	ACCAGCATCTGTGATCTCT	CCGGCTGAGTCAATAGACT	TTCTCTTCACTCAACCATG	5100
	CAAGAAAGCTGTCTTAACAT	CAAGTTGAGCTAGCTCCAAA	TTCAACTGTGCTACCAAGGC	CAACAAAATTTAATGGAGG	AATGCTTCAACACAGGGGAA	5200
	AATACATCATTTGATGCAAT	TCCCTCCTTCTGAGCGTAGC	CTTTAGCTACCAATCTTGCC	TTGTAGCGAATATCTTCTT	GCTAGGAGATC	5291
	20	40	60	80	100	

Figure 1. Nucleotide and deduced amino acid sequences of the cotton *Mat5-A* allele and the *Mat5-D* allele cDNAs encoding methionine-rich 2S albumin storage proteins. The sequence of *Mat5-A* clone GC164-24RC is shown as the reference. Only where the sequences of the cDNAs are different from those of *Mat5-A* are the cDNA sequences shown above those of *Mat5-A*. The probable TATAA sequence and polyadenylation addition sequence are both shown in bold. Pairs of long direct repeats are underlined and numbered below their 5' end. The limits of a repetitive element found in an unrelated cotton *Lea4* gene are indicated by underlines at nucleotides 907 and 1795. Sequences similar to those found in other 2S albumin storage protein genes (8, 9) are indicated by double underlines.

Table 1. Characteristics of the *Mat5-A* gene and *Mat5-D* cDNAs from *Gossypium hirsutum*

Organism:
Gossypium hirsutum L cv Coker 201 (Upland cotton), Malvaceae.

Function:
2S albumin storage protein.

Expression:
During the maturation stage in embryo development; coordinately expressed with many other *Mat* genes, including those of vicilin and legumin storage proteins (5, 6).

Source:
Nuclear DNA from embryo cotyledons 20 to 23 d postanthesis (preendoreduplication). Partial *Sau3A*I digest was cloned in λ GEM-12 (Promega), and three phage were identified by hybridization with cDNA clone C164. Clone C164 has been described (5).

Genome:
Restriction fragment and sequence analysis indicates that the three phage isolates are from the *Mat5-A* allele in the A genome (G.A. Galau, unpublished data). The extensive sequence difference between *Mat5-A* and the sequenced cDNAs indicates that the cDNAs must be transcribed from the other allele, *Mat5-D*.

Sequence of *Mat5-A* GC164-24RC:
Cloning of the 7.2-kilobase *EcoRI*/polylinker *EcoRI* fragment of phage isolate 24 into Bluescript (Stratagene), subcloning of a 5.3-kilobase *Clal*/polylinker *EcoRI* fragment, followed by dideoxy sequencing of both strands with *Taq* polymerase using cloned restriction fragments and unidirectional deletions as double-stranded templates. The polylinker sequence at its 3' end is not reported. Where checked in the transcribed region, the sequences of isolates GC164-45 and GC164-27 are identical with that of GC164-24RC.

Sequence of *Mat5-D* cDNA C164:
The entire cDNA was sequenced on both strands using cloned restriction fragments as templates. The clone contains an artifact at its 5' end: a 250-nucleotide-long perfect inverse repeat of a part of its 3' end (C164 nucleotide 307-556). The artifact is not reported.

Partial Sequence of Other *Mat5-D* cDNA Clones:
Three primers (leftward *Mat5-A* nucleotide 3268-3248 and rightward *Mat5-A* nucleotide 3198-3198 and 3555-3572) were used to sequence most of each of the clones C42, C87, C90, C103, and C115 on at least one strand. Clones C103 and C115 contain artifacts similar to that in C164.

Transcription Start:
The leftward primer *Mat5-A* nucleotide 3268-3248 was used to determine, by primer extension, the transcription start during the maturation stage.

Genbank Accession Nos.:
M86213, *Mat5-A* GC164-24RC; M83301, *Mat5-A* cDNA C164.

	Signal Peptide	Discard	Small Subunit	Discard
Gossypium	<u>MA</u> - <u>KL</u> - <u>AVYLATL</u> - <u>ALI</u> - <u>LFLANA</u>	<u>S</u> <u>I</u> <u>T</u> <u>S</u> <u>V</u> - <u>IV</u> <u>E</u> <u>S</u> <u>E</u> <u>E</u> - <u>---</u> <u>N</u>	<u>---</u> <u>RD</u> <u>S</u> <u>C</u> - <u>---</u> <u>EQ</u> <u>Q</u> <u>I</u> <u>R</u> <u>K</u> <u>Q</u> <u>A</u> <u>H</u> <u>L</u> <u>K</u> <u>H</u> <u>C</u> <u>Q</u> <u>K</u> <u>Y</u> <u>M</u> <u>E</u> - <u>---</u> <u>E</u> <u>E</u> <u>L</u> - <u>GG</u>	<u>EG</u> <u>S</u> - <u>---</u> <u>D</u> - <u>---</u>
Bertholletia	<u>MA</u> - <u>KI</u> - <u>SV</u> <u>AA</u> <u>AA</u> <u>LL</u> <u>V</u> <u>L</u> <u>M</u> <u>AL</u> <u>GH</u> <u>ATA</u>	<u>F</u> <u>R</u> <u>A</u> <u>T</u> <u>V</u> <u>T</u> <u>T</u> <u>T</u> <u>V</u> <u>V</u> <u>E</u> <u>E</u> <u>E</u> - <u>---</u> <u>N</u>	<u>---</u> <u>Q</u> <u>E</u> <u>E</u> <u>C</u> <u>R</u> - <u>EQ</u> <u>M</u> <u>Q</u> <u>R</u> - <u>Q</u> <u>Q</u> <u>M</u> <u>L</u> <u>S</u> <u>H</u> <u>C</u> <u>R</u> <u>M</u> <u>Y</u> <u>M</u> <u>R</u> <u>Q</u> <u>Q</u> <u>M</u> <u>E</u> <u>S</u> - <u>---</u>	<u>-P</u> <u>Y</u> <u>Q</u> <u>T</u> <u>M</u> - <u>---</u>
Arabidopsis	<u>M</u> <u>A</u> <u>N</u> <u>K</u> <u>L</u> <u>F</u> <u>L</u> <u>V</u> <u>C</u> - <u>AA</u> <u>L</u> - <u>AL</u> <u>C</u> <u>F</u> <u>L</u> - <u>LT</u> <u>NA</u>	<u>S</u> <u>I</u> <u>Y</u> <u>R</u> - <u>---</u> <u>IV</u> <u>V</u> <u>E</u> <u>F</u> <u>E</u> <u>E</u> <u>D</u> <u>D</u> <u>A</u> <u>S</u> <u>N</u>	<u>P</u> <u>I</u> <u>G</u> <u>P</u> <u>R</u> <u>Q</u> <u>K</u> <u>C</u> <u>R</u> <u>K</u> <u>E</u> <u>F</u> <u>Q</u> <u>Q</u> - <u>---</u> <u>S</u> <u>Q</u> <u>H</u> <u>L</u> <u>R</u> <u>A</u> <u>C</u> <u>Q</u> <u>K</u> <u>L</u> <u>M</u> <u>R</u> <u>K</u> <u>Q</u> <u>M</u> <u>R</u> <u>Q</u> <u>G</u> <u>R</u> <u>G</u> <u>G</u>	<u>G</u> <u>P</u> <u>S</u> <u>L</u> <u>D</u> <u>D</u> <u>E</u> <u>F</u> <u>D</u>
	Large Subunit			Discard
Gossypium	<u>NI</u> <u>A</u> <u>G</u> - <u>---</u> <u>GY</u> <u>I</u> <u>D</u> <u>S</u> - <u>---</u> <u>CC</u> - <u>Q</u> <u>L</u> <u>E</u> <u>K</u> <u>M</u> <u>D</u> <u>T</u> <u>Q</u> - <u>C</u> <u>R</u> <u>C</u> <u>Q</u> <u>G</u> <u>L</u> <u>R</u> <u>H</u> <u>A</u> <u>T</u> <u>M</u> <u>Q</u> <u>Q</u> <u>M</u> <u>Q</u> <u>Q</u> <u>M</u> <u>Q</u> <u>Q</u> <u>M</u> - <u>G</u> <u>S</u> <u>K</u> <u>Q</u> <u>M</u> <u>R</u> <u>E</u> <u>I</u> <u>M</u> <u>Q</u> <u>K</u> <u>V</u> <u>I</u> <u>K</u> <u>K</u> - <u>I</u> <u>M</u> <u>S</u> <u>E</u> <u>C</u> <u>E</u> <u>M</u> <u>E</u> <u>P</u> <u>G</u> <u>R</u> - <u>---</u> <u>CD</u> <u>T</u> <u>P</u> <u>S</u> <u>R</u> <u>S</u> <u>L</u>			<u>I</u>
Bertholletia	<u>P</u> <u>R</u> <u>R</u> <u>G</u> - <u>M</u> <u>E</u> <u>P</u> <u>H</u> <u>M</u> <u>S</u> <u>E</u> - <u>---</u> <u>CC</u> - <u>E</u> <u>Q</u> <u>L</u> <u>E</u> <u>G</u> <u>M</u> <u>D</u> <u>E</u> <u>S</u> - <u>C</u> <u>R</u> <u>C</u> <u>E</u> <u>G</u> <u>L</u> <u>R</u> <u>M</u> <u>M</u> <u>M</u> <u>R</u> - <u>M</u> <u>Q</u> <u>Q</u> <u>E</u> <u>E</u> <u>M</u> <u>Q</u> <u>P</u> <u>R</u> <u>G</u> - <u>E</u> <u>Q</u> <u>M</u> <u>R</u> <u>R</u> <u>M</u> <u>M</u> <u>R</u> - <u>---</u> <u>LA</u> <u>E</u> <u>N</u> <u>I</u> <u>P</u> <u>S</u> <u>R</u> <u>C</u> <u>N</u> <u>L</u> <u>S</u> <u>P</u> <u>M</u> <u>R</u> - <u>---</u> <u>C</u> - <u>---</u> <u>P</u> <u>M</u> <u>G</u> <u>G</u> <u>S</u>			<u>I</u> <u>A</u> <u>G</u> <u>F</u>
Arabidopsis	<u>P</u> - <u>Q</u> <u>G</u> <u>P</u> <u>Q</u> <u>G</u> <u>R</u> <u>P</u> <u>Q</u> <u>L</u> <u>L</u> <u>Q</u> <u>Q</u> <u>C</u> <u>M</u> <u>E</u> - <u>---</u> <u>LR</u> <u>Q</u> - <u>E</u> <u>E</u> <u>P</u> <u>V</u> <u>C</u> <u>V</u> <u>C</u> <u>P</u> <u>T</u> <u>L</u> <u>R</u> <u>Q</u> <u>A</u> <u>A</u> <u>K</u> - <u>---</u> <u>AV</u> <u>S</u> <u>L</u> <u>Q</u> <u>G</u> <u>Q</u> <u>H</u> <u>G</u> <u>P</u> <u>E</u> <u>Q</u> <u>V</u> <u>R</u> <u>K</u> <u>I</u> <u>Y</u> <u>Q</u> - <u>---</u> <u>T</u> <u>A</u> <u>K</u> <u>Y</u> <u>L</u> <u>P</u> <u>N</u> <u>I</u> <u>C</u> <u>K</u> <u>I</u> - <u>P</u> <u>Q</u> <u>V</u> <u>G</u> <u>V</u> <u>C</u> - <u>---</u> <u>P</u> <u>F</u> <u>Q</u> <u>T</u>			<u>I</u> <u>P</u> <u>F</u> <u>F</u> <u>P</u> <u>S</u>

Figure 2. Alignment of amino acid sequences of *Gossypium*, *Bertholletia*, and *Arabidopsis* 2S albumin storage proteins. The sequences are *Mat5-A* from *Gossypium*, pHS-3 from *Bertholletia* (1), and a composite of four *Arabidopsis* sequences (9), giving weight to residues that are held in common with either of the other two sequences. The processing sites are those reported in *Arabidopsis* (9). Residues are underlined if they are shared in at least two of the proteins.

LITERATURE CITED

1. Altenbach SB, Pearson KW, Leung FW, Sun SSM (1987) Cloning and sequence analysis of a cDNA encoding a Brazil nut protein exceptionally rich in methionine. *Plant Mol Biol* **8**: 239–250
2. Chlan CA, Borroto K, Kamalay JA, Dure L III (1987) Developmental biochemistry of cottonseed embryogenesis and germination. XIX. Sequences and genomic organization of the α globulin (vicilin) genes of cottonseed. *Plant Mol Biol* **9**: 533–546
3. Galau GA, Wang HY-C, Hughes DW (1991) Sequence of the *Gossypium hirsutum* D-genome allele of *Legumin A* and its mRNA. *Plant Physiol* **97**: 1268–1270
4. Galau GA, Wang HY-C, Hughes DW (1992) Cotton *Lea4* (D19) and *Lea42* (D132) Group 1 *Lea* genes encoding water stress-related proteins containing a 20-amino acid motif. *Plant Physiol* **99**: 738–788
5. Goldberg RB, Barker SJ, Perez-Grau L (1989) Regulation of gene expression during plant embryogenesis. *Cell* **56**: 149–160
6. Hughes DW, Galau GA (1989) Temporally modular gene expression during cotyledon development. *Genes Dev* **3**: 358–369
7. Hughes DW, Galau GA (1991) Developmental and environmental induction of *Lea* and *LeaA* mRNAs and the postabscission program during embryo culture. *Plant Cell* **3**: 605–618
8. Irwin SD, Keen JN, Findlay JBC, Lord JM (1990) The *Ricinus communis* 2S albumin precursor: a single preproprotein may be processed into two different heterodimeric storage proteins. *Mol Gen Genet* **222**: 400–408
9. Krebbers E, Herdies L, De Clercq A, Seurink J, Leemans J, VanDamme J, Segura M, Gheysen G, Van Montagu M, Vandekerckhove J (1988) Determination of the processing sites of an *Arabidopsis* 2S albumin and characterization of the complete gene family. *Plant Physiol* **87**: 859–866
10. Voytas DF, Ausubel FM (1988) A copia-like transposable element family in *Arabidopsis thaliana*. *Nature* **336**: 242–244