

Cloning and sequence analysis of a cDNA encoding a Brazil nut protein exceptionally rich in methionine

Susan B. Altenbach, Karen W. Pearson, Filomena W. Leung & Samuel S. M. Sun*
ARCO Plant Cell Research Institute, 6560 Trinity Court, Dublin, CA 94568, USA (*for offprints)

Received 25 September 1986; accepted 28 October 1986

Keywords: Brazil nut, methionine-rich protein, oilseed proteins, seed storage proteins

Abstract

The primary amino acid sequence of an abundant methionine-rich seed protein found in Brazil nut (*Bertholletia excelsa* H.B.K.) has been elucidated by protein sequencing and from the nucleotide sequence of cDNA clones. The 9 kDa subunit of this protein was found to contain 77 amino acids of which 14 were methionine (18%) and 6 were cysteine (8%). Over half of the methionine residues in this subunit are clustered in two regions of the polypeptide where they are interspersed with arginine residues. In one of these regions, methionine residues account for 5 out of 6 amino acids and four of these methionine residues are contiguous. The sequence data verifies that the Brazil nut sulfur-rich protein is synthesized as a precursor polypeptide that is considerably larger than either of the two subunits of the mature protein. Three proteolytic processing steps by which the encoded polypeptide is sequentially trimmed to the 9 kDa and 3 kDa subunit polypeptides have been correlated with the sequence information. In addition, we have found that the sulfur-rich protein from Brazil nut is homologous in its amino acid sequence to small water-soluble proteins found in two other oilseeds, castor bean (*Ricinus communis*) and rapeseed (*Brassica napus*). When the amino acid sequences of these three proteins are aligned to maximize homology, the arrangement of cysteine residues is conserved. However, the two subunits of the Brazil nut protein contain over 19% methionine whereas the homologous proteins from castor bean and rapeseed contain only 2.1% and 2.6% methionine, respectively.

Introduction

In contrast to the seed proteins from many plants which contain relatively low levels of the sulfur-containing amino acids, the seed proteins from Brazil nut (*Bertholletia excelsa* H.B.K.) contain large percentages of methionine and cysteine, 8.3%–9.1% by weight [3, 26]. From a 2S albumin fraction of Brazil nut proteins, we previously purified an abundant sulfur-rich protein. This sulfur-rich protein consists of two low molecular weight subunits, a 9 kDa polypeptide and a 3 kDa polypeptide, which associate through disulfide bridges

to form a 12 kDa protein molecule (unpublished data). The sulfur-rich protein is synthesized in the seed only at a particular developmental stage, about 8 to 9 months after flowering. *In vitro* and *in vivo* labelling studies have indicated that this protein is synthesized initially as a larger precursor polypeptide of about 18 kDa which then undergoes three proteolytic processing steps before it attains its mature form [2].

We now report the amino acid sequence of some 77% of the large subunit of the sulfur-rich protein obtained by Edman degradation. Using a synthetic oligodeoxynucleotide probe whose sequence was

based on a methionine-rich region found in this partial amino acid sequence, we have identified cDNA clones encoding the sulfur-rich protein. In this paper, we present the complete nucleotide sequence of one Brazil nut cDNA clone and verify that the sulfur-rich protein encoded by this clone is synthesized as a larger precursor polypeptide. We have correlated the three processing steps by which the encoded polypeptide is sequentially trimmed to the 9 kDa and 3 kDa polypeptides with the sequence information and demonstrate that the 9 kDa subunit encoded by this clone contains 18% methionine and 8% cysteine. Finally, a computer search of available protein sequences revealed that the methionine-rich protein from Brazil nut is homologous in its amino acid sequence to small water-soluble seed proteins found in castor bean and rapeseed which contain only modest levels of methionine.

Materials and methods

Plant material

Brazil nuts are indigenous to the Amazon River basin; they do not grow anywhere in the United States. Brazil nut fruits were obtained approximately 9 months after flowering from Brazil (Manaus) or Peru (Iquitos or Puerto Maldonado).

Purification of the sulfur-rich protein and amino acid sequence determination

Brazil nut embryos were ground into a fine paste and defatted by extraction with hexane. The resulting defatted Brazil nut flour was then extracted in a buffer containing 1 M NaCl in 0.035 M sodium phosphate buffer, pH 7.5. The sulfur-rich protein was purified from this crude extract by the procedure of Youle and Huang [26]. The resulting sucrose gradient fractions were dialyzed extensively against deionized water at 4°C to precipitate the contaminating globulin proteins. The final protein sample contained polypeptides of 9 kDa and 3 kDa when analyzed on SDS-20% polyacrylamide gels.

The protein sample for sequencing was prepared by incubation of 2 mg of the purified sulfur-rich protein with 1 M Tris-HCl buffer, pH 8.8, containing 1 mM EDTA and 0.15 M 2-mercaptoethanol at 37°C under nitrogen gas for 4.5 hours. At the end of the incubation, iodoacetic acid was added to a final concentration of 0.22 M and the sample was incubated at 37°C in the dark for 30 minutes. After this treatment, the protein sample was dialyzed extensively against deionized water and lyophilized.

Sequence analysis of the sulfur-rich protein was performed by automated Edman degradation [6] on a Beckman 890C liquid-phase sequenator equipped with a cold trap using program 050783 with 0.1 M Quadrol (Beckman Instruments, Inc.) and polybrene (2 mg) as a carrier. About 10 nmol of the sulfur-rich protein were applied into the liquid phase sequenator. Norleucine was added to the fractions and used as an internal standard for quantitation of each cycle. Phenylthiohydantoin-amino acids were identified and measured by (Packard 419) gas liquid chromatography [20], (Water 6000A) high performance liquid chromatography [4] and thin layer chromatography [9]. At least two of these methods were used at each step. A total of 60 cycles of degradation were conducted, and 97% repetitive yield was observed. No PTH-amino acid could be identified after cycle 57.

Preparation of cDNA library and isolation of clones

Polyadenylated RNA was prepared from 9-month-old developing Brazil nut seeds by methods described previously for *Phaseolus vulgaris* [7] and was cloned in the dimer-primer vector pARC7 [1]. The resulting clones were screened by colony hybridization [24] using a 5' labelled probe which consisted of a mixture of 6 synthetic oligodeoxynucleotides complementary to the 6 possible RNA sequences which could encode a methionine-rich region found in the partial amino acid sequence of the 9 kDa subunit of the sulfur-rich protein (Fig. 1B). The probe was hybridized to the filters in 6× NET (0.9 M NaCl, 0.09 M Tris/Cl, pH 7.5, 0.006 M EDTA), 0.05% NP-40, and 250 µg/ml yeast tRNA at 37°C for 20 hours. The filters were

washed in $6 \times \text{SSC}$ at 37°C before autoradiography.

Sequencing of cDNAs and primer extension analysis

The sequence of cDNA clone pHS-3 was determined from both DNA strands by the dideoxy chain termination method [22]. Where necessary, regions of the clone were also sequenced by the method of Maxam and Gilbert [13]. The 25 nucleotides at the 5' end of the mRNA encoding the sulfur-rich protein were not represented in pHS-3 but were obtained by using a synthetic oligodeoxynucleotide complementary to nucleotides #49–69 as a primer to synthesize DNA complementary to the 5' end of the mRNA and sequencing the resulting extension product. For primer extension, the oligodeoxynucleotide 5' AATCTTCGCCATGGT-GATTCT 3', labelled at its 5' end, was annealed to $3 \mu\text{g}$ of poly(A)⁺ RNA from the seeds of 9-month-old Brazil nuts in 8 mM Tris pH 7.5, 5 mM EDTA at 90°C for 5 minutes. NaCl was added to 0.1 M and the sample was incubated for 20 minutes at 90°C followed by 15 minutes at 25°C . The annealed DNA sample was brought to a final concentration of 50 mM Tris pH 8.3, 5 mM DTT, 15 mM MgCl_2 , 0.5 mM dNTPs, and $0.1 \mu\text{g/ml}$ BSA. AMV reverse transcriptase (BRL, 37.5 units) was added and the reaction was incubated at 37°C for 90 minutes. EDTA was added to 20 mM and the sample was extracted twice with phenol:chloroform:isoamyl alcohol (25:24:1) and precipitated with ethanol. After denaturation, the samples were subjected to electrophoresis on an 8% sequencing gel. Three bands resulted which differed in length by single nucleotides. DNA from each of the three bands was eluted from the gel and sequenced by the method of Maxam and Gilbert [13].

Hybrid-selected translation of cDNA clones

Characterization of cDNA clones by hybrid-selected translation was performed as described by Maniatis [12]. Three micrograms of either pHS-3 or

pARC-7 plasmid DNA were denatured, bound to nitrocellulose paper and hybridized to $2 \mu\text{g}$ of Poly(A)⁺ RNA prepared from 9-month-old Brazil nut seeds. RNA which was specifically bound to the DNA was then eluted, precipitated with ethanol along with $5 \mu\text{g}$ carrier yeast tRNA, and translated in a wheat germ system [7]. In addition, the translation products directed by RNA selected by pHS-3 were immunoprecipitated [7] with a polyclonal antibody which had been made to a mixture of the 9 kDa and 3 kDa components of the mature Brazil nut sulfur-rich protein and its 12 kDa precursor. The proteins, labelled with [³⁵S]methionine, were analyzed on a SDS-20% polyacrylamide gel [11] and visualized by autoradiography.

Results

Partial amino acid sequence of the sulfur-rich protein

Two amino acid sequences were obtained from the analysis of the carboxymethylated sulfur-rich protein; one major (80%) and one minor (20%). The major sequence starts with Pro–Arg–Arg–Gly–Met... as NH_2 -terminal amino acids, while the minor one starts with Gly–Met... (Fig. 1A). The two protein sequences are identical in the region sequenced except that the minor one is three amino acids shorter than the major one at the NH_2 terminus; thus we were able to determine the first 57 amino acids for the major sequence and the first 54 amino acids for the minor one. The sulfur-rich protein consists of two subunits, a 9 kDa polypeptide and 3 kDa polypeptide. Both the 54 and the 57 amino acid sequences exceed the length of the 3 kDa polypeptide, thus these sequences must represent 9 kDa polypeptides, possibly two members of the 9 kDa polypeptide family. We did not obtain any amino acid sequence for the 3 kDa polypeptide, suggesting that either the 3 kDa polypeptide sequence is identical to the 9 kDa sequence or the NH_2 terminus of the 3 kDa polypeptide is blocked (see Discussion).

This amino acid sequence represents about 77% of the 9 kDa subunit. The sequence contains un-

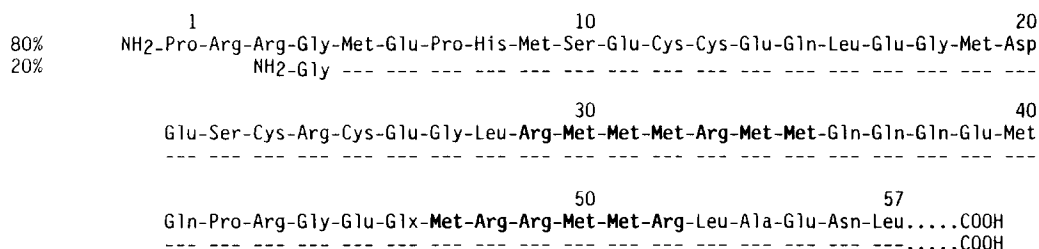
A

Fig. 1A. The partial amino acid sequence of the 9 kDa subunit of the sulfur-rich protein from Brazil nut. After reduction and carboxy-methylation, the purified sulfur-rich protein was sequenced using an automatic liquid-phase sequenator. Two sequences, one major (80%) with Pro as the NH₂ terminal amino acid (shown in the first line), and one minor (20%) with Gly as the NH₂ terminal amino acid (shown in the second line), were detected. Amino acid residues found in the minor sequence which are identical to those found in the major sequence are indicated with dashes. Methionine-rich regions found in the partial sequence are highlighted.

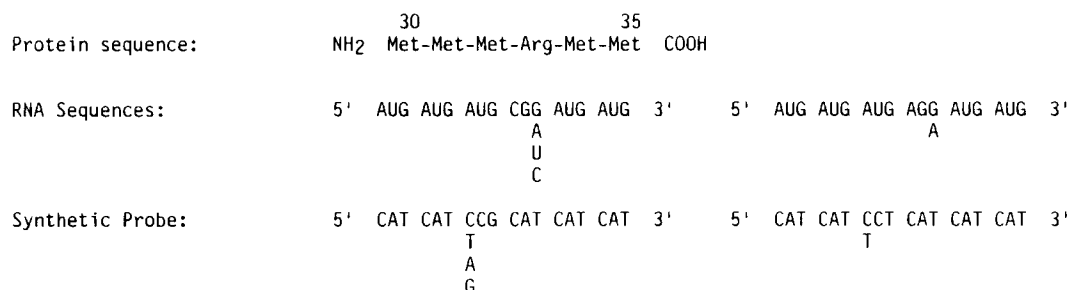
B

Fig. 1B. Amino acid sequence of the first methionine-rich region which was used as a basis for a synthetic oligodeoxynucleotide probe. The sequences of the 6 possible mRNAs encoding this portion of the protein sequence are shown in the second line and the sequences included in the synthetic oligodeoxynucleotide probe complementary to the mRNA are shown in the bottom line.

usually high levels of the sulfur amino acids: 21% methionine and 7% cysteine. There are two regions in the partial amino acid sequence where methionine residues are clustered with arginine residues: residues #29–35 (*Arg–Met–Met–Met–Arg–Met–Met*) and residues #47–52 (*Met–Arg–Arg–Met–Met–Arg*) (Fig. 1A).

Identification and characterization of cDNA clones encoding the sulfur-rich protein

An oligodeoxynucleotide probe was synthesized (by Bioscience, Inc.) which was complementary to the 6 possible RNA sequences encoding one of these methionine-rich regions (amino acid residues

#30–35) (Fig. 1B). This oligodeoxynucleotide probe hybridized to a number of clones from a cDNA library prepared using RNA from 9-month-old Brazil nut seeds. Twelve of these clones with inserts ranging from 350 bp to 700 bp were selected for further analysis.

Sequence analysis of one of these clones, pHS-3, demonstrates unequivocally that this cDNA encodes a polypeptide which is extremely rich in the sulfur-containing amino acids (Fig. 2A). The sequence of pHS-3 is 599 nucleotides long excluding the poly(A) tail. By primer extension analysis using a 21 base synthetic oligodeoxynucleotide complementary to a region near the 5' end of pHS-3, we determined that this cDNA clone falls 25 nucleotides short of the 5' end of the mRNA en-

A

5'	NNN TCC ACC ACT GCT CTG TAT CAC ATA TAC CAA CCC ACG CTT ATA CCC AGA ATC ACC	17K → ATG Met	60 1
	GCG AAG ATT TCA GTT GCG GCA GCA GCC CTC CTT GTC CTC ATG GCC CTC GGC CAC GCC ACC GCC TTC	15K →	126 23
	Ala Lys Ile Ser Val Ala Ala Ala Ala Leu Leu Val Leu Met Ala Leu Gly His Ala Thr Ala Phe		
	CGG GCC ACC GTC ACC ACC ACA GTG GTG GAG GAG GAG AAC CAG GAG GAG TGT CGC GAG CAG ATG CAG		192 45
	Arg Ala Thr Val Thr Thr Thr Val Val Glu Glu Glu Asn Gln Glu Glu Cys Arg Glu Gln Met Gln		
	12K → AGA CAG CAG ATG CTC AGC CAC TGC CGG ATG TAC ATG AGA CAG CAG ATG GAG GAG AGC CCG TAC CAG		258 67
	Arg Gln Gln Met Leu Ser His Cys Arg Met Tyr Met Arg Gln Gln Met Glu Glu Ser Pro Tyr Gln		
	9K(80%) → ACC ATG CCC AGG CGG GGA ATG GAG CCG CAC ATG AGC GAG TGC TGC GAG CAG CTG GAG GGG ATG GAC	9K(20%) →	324 89
	Thr Met Pro Arg Arg Gly Met Glu Pro His Met Ser Glu Cys Cys Glu Gln Leu Glu Gly Met Asp		
	Pro Arg Arg Gly Met Glu Pro His Met Ser Glu Cys Cys Glu Gln Leu Glu Gly Met Asp		
	GAG AGC TGC AGA TGC GAA GGC TTA AGG ATG ATG ATG ATG AGG ATG CAA CAG GAG GAG ATG CAA CCC		390 111
	Glu Ser Cys Arg Cys Glu Gly Leu Arg Met Met Met Met Arg Met Gln Gln Glu Glu Met Gln Pro		
	Glu Ser Cys Arg Cys Glu Gly Leu Arg Met Met Met Arg Met Met Gln Gln Gln Glu Met Gln Pro		
	CGA GGG GAG CAG ATG CGA AGG ATG ATG AGG CTG GCC GAG AAT ATC CCT TCC CGC TGC AAC CTC AGT		456 133
	Arg Gly Glu Gln Met Arg Arg Met Met Arg Leu Ala Glu Asn Ile Pro Ser Arg Cys Asn Leu Ser		
	Arg Gly Glu Glu Met Arg Arg Met Met Arg Leu Ala Glu Asn Leu		
	CCC ATG AGA TGC CCC ATG GGT GGC TCC ATT GCC GGG TTC TGA ATC TGC CAC TAG CCA GTG CTG TAA		522 146
	Pro Met Arg Cys Pro Met Gly Gly Ser Ile Ala Gly Phe		
	ATG TTA ATA AGG CTC TCA CAA ACT AGC TCT TTG TTG GCT TTT GGC CGG AGA CTA GGG TGT GGG ATT		588
	AAT AAT AAT AGC ACA CTA TCG TGT GTT CTC AGC TTC (A) _n		624

Fig. 2A. The complete nucleotide sequence of a cDNA which encodes the sulfur-rich protein from Brazil nut. The sequence of nucleotides #26 to 624 as well as the poly(A) tract was determined from the analysis of the cDNA clone pHS-3. The first 25 bases were obtained by sequencing the largest primer extension product which was synthesized using a synthetic oligodeoxynucleotide complementary to residues #49–69 as a primer. The first three nucleotides were uncertain from sequencing and are represented by NNN in the sequence. The first ATG codon from the 5' end (nucleotides #57–59) and the termination codon TGA (nucleotides #495–497) are marked in boxes. The amino acid sequence deduced from the nucleotide sequence of the resulting open reading frame is shown in the second line and the major 57 residue partial amino acid sequence which was determined from analysis of the purified sulfur-rich protein is shown in line 3. The approximate sites of cleavages which may be involved in the maturation of the sulfur-rich protein are marked with arrows above the nucleotide sequence. ATG codons in the DNA sequence as well as methionine residues in the protein sequence are highlighted. Numbers in the right margins refer to the number of nucleotides or amino acids.

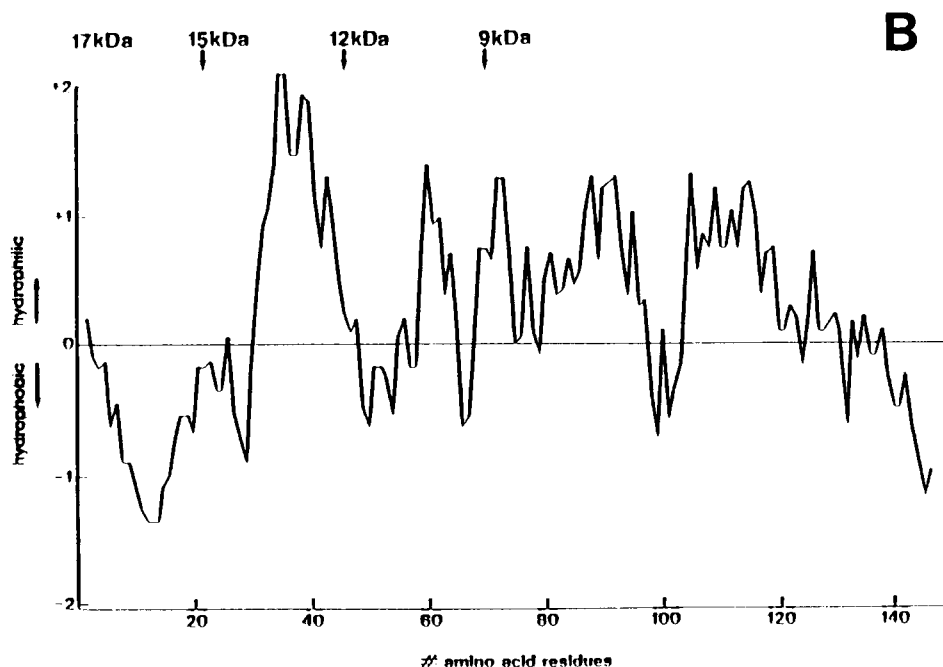


Fig. 2B. Hydropathy plot of Brazil nut sulfur-rich protein encoded by cDNA clone pHS-3. Plot showing the hydrophobic and hydrophilic regions of the protein encoded by pHS-3 was generated using the algorithm of Hopp and Wood [8]. The horizontal line in the middle of the plot represents a hydrophilicity value of 0. Hydrophilic regions are plotted above the 0-line and hydrophobic regions beneath the 0-line. Numbers along the x-axis refer to the number of amino acid residues from the NH_2 terminus of the protein. The approximate location of cleavages believed to be involved in the maturation of the Brazil nut sulfur-rich protein are shown with arrows.

coding the sulfur-rich protein. There are no ATG codons in the sequence of the primer extension product; thus, the first ATG codon found in pHS-3 (residues #58–60) represents the initiation codon for protein synthesis. This ATG fits with the consensus sequence for eucaryotic protein initiation sites [10]. A stop codon, TGA, is encoded by nucleotides #495–497 in pHS-3. The resulting open reading frame could encode a polypeptide of 146 amino acids, of which over 20% are sulfur-containing amino acids; 15.1% of these residues are methionine while 5.5% are cysteine. The first portion of the encoded polypeptide contains a large proportion of hydrophobic residues; of the 22 residues at the amino terminus of the protein, 36% are alanine and 18% are leucine. In comparison, the rest of the polypeptide is rich in arginine, glutamine and glutamic acid, a composition which is characteristic of other plant seed storage proteins. A hydropathy plot (Fig. 2B) demonstrates that the amino terminus of the polypeptide is hydrophobic

while the remainder of the polypeptide is largely hydrophilic.

By aligning the amino acid sequence derived from the nucleotide sequence with the major sequence determined from the purified 9 kDa subunit, we have found that the coding region for the 9 kDa polypeptide begins 265 nucleotides from the 5' end of the mRNA. By adding up the molecular weights of the individual amino acids encoded by this region, we arrive at a value of almost 9 kDa. The amino acid sequence derived from the nucleotide sequence of the portion of the open reading frame between nucleotides 265 and 435 agrees quite well, although not precisely, with the major 57 residue partial amino acid sequence of the 9 kDa subunit of the sulfur-rich protein (Fig. 2A). Methionine residues are very predominant in the 9 kDa subunit of the mature protein. There are 14 methionine residues in this region, representing 18.2% of the 77 amino acid polypeptide. Eight of these 14 methionines are found clustered with

arginine residues in two regions of the polypeptide. In the first cluster, between amino acid residues #99 and 104, five out of six residues are methionines and four of the methionine residues are contiguous. The second methionine cluster, between amino acid residues #116 and 121, includes three methionine residues and three arginine residues. Interestingly, 2 of the 4 amino acid differences which are found between the amino acid sequence determined from the protein and that derived from the nucleotide sequence are found in the methionine-rich region that was used as a basis for the synthetic oligodeoxynucleotide probe. A second cDNA clone selected by the same probe was perfectly homologous with one of the sequences represented in the probe (unpublished data), suggesting that the sulfur-rich protein is encoded by a family of genes with some variation in these methionine-rich regions. The 9 kDa subunit of the Brazil nut protein also contains a high proportion of cysteine (7.7%).

By hybrid-selected translation, we have found that pHS-3 is able to select a mRNA from a population of 9-month-old Brazil nut RNAs which directs the synthesis of an 18 kDa polypeptide *in vitro* (Fig. 3). This 18 kDa polypeptide is immunoprecipitable with a polyclonal antibody raised in rabbits against the purified Brazil nut sulfur-rich protein, demonstrating conclusively that the sulfur-rich protein is synthesized initially as a larger precursor polypeptide.

Homology of the sulfur-rich protein to other water-soluble seed proteins

In a computer search of proteins whose amino acid sequences have been determined, we found that the sulfur-rich protein from Brazil nut shares a great deal of homology with both the large and the small subunits of a low molecular weight and water-soluble seed storage protein from castor bean (*Ricinus communis*) [23]. We have aligned the amino acid sequence of the small subunit of this castor bean protein with the Brazil nut sequence starting at amino acid residue #35 and that of the large subunit of the castor bean protein with amino acid

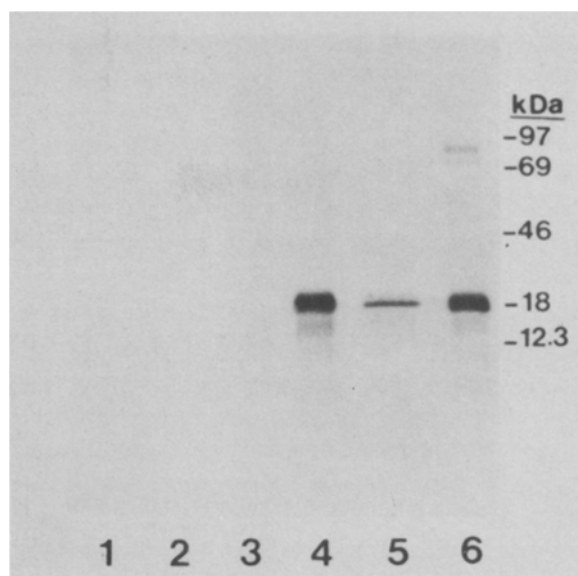


Fig. 3. Identification of a cDNA clone for the sulfur-rich Brazil nut protein by translation of hybrid-selected mRNAs. Lane 1 shows endogenous proteins synthesized in the wheat germ system and labelled with [^{35} S] methionine. Lanes 2 and 3 show labelled proteins synthesized by either RNA selected by the vector pARC7 or 5 μ g yeast tRNA. The translation products of RNA selected by pHS-3 are displayed in lane 4 and these products are immunoprecipitated with the Brazil nut sulfur-rich protein antibody in lane 5. Lane 6 shows the total translation products of Brazil nut poly(A) $^{+}$ RNA in the wheat germ system.

residues starting at #72 (Fig. 4A). Allowing 2 small gaps in the small subunit comparison and 4 small gaps in the large subunit comparison to maximize sequence homology, we find over 44% homology between the castor bean protein and the Brazil nut sulfur-rich protein. Both proteins are high in glutamine/glutamic acid and arginine (22% and 13% for the Brazil nut protein and 29.5% and 10.5% for the castor bean protein, respectively), and the positions of many of these residues are conserved in the two proteins. Interestingly, both the Brazil nut and the castor bean proteins are relatively rich in cysteine (7% and 8.4%, respectively) and the positions of these residues are similar in both proteins. Another small water-soluble protein found in rapeseed (*Brassica napus*), napin [5], shows some homology (about 21%) with the Brazil nut protein (Fig. 4A).

A

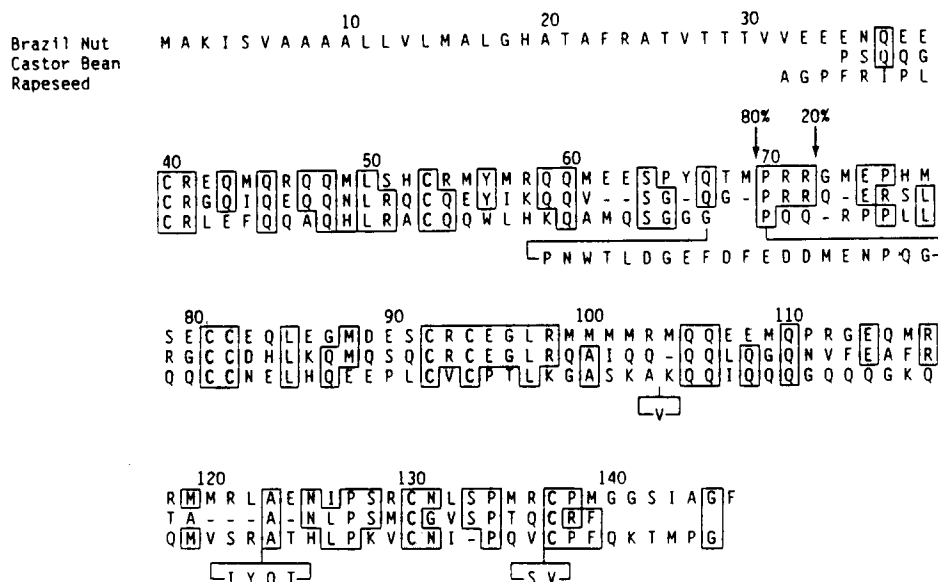


Fig. 4A. Comparison of the amino acid sequences of the 2S sulfur-rich protein of Brazil nut and the 2S water-soluble seed proteins from castor bean and rapeseed. The top line shows the amino acid sequence of the precursor polypeptide for the Brazil nut sulfur-rich protein (derived from the nucleotide sequence of cDNA clone pHS-3). The amino acid sequences of both subunits of the castor bean protein [23] are shown in the second line and the amino acid sequences of both subunits of napin (derived from the sequences of napin cDNA clones) [5] are shown in the third line. The one letter amino acid code was used for these comparisons and the sequences were aligned to maximize homology with the Brazil nut protein. Homologous residues are enclosed in boxes. The locations of the final processing sites involved in the maturation of the Brazil nut sulfur-rich protein (major, 80%, and minor, 20%) were determined from the amino acid sequence analysis of the 9 kDa subunit and are indicated with arrows.

Although the homology noted between Brazil nut and rapeseed proteins is substantially less than that between the Brazil nut and castor bean proteins, 24 out of 28 amino acids (85.7%) conserved between the Brazil nut and rapeseed proteins are also common to the castor bean protein. In addition, the positions of cysteine residues in all three proteins are conserved. However, the Brazil nut protein is unusually rich in methionine (19%) while the castor bean and rapeseed proteins contain only about 2% methionine. Thus, a large percentage of the non-homology between the Brazil nut protein and the castor bean or rapeseed protein sequences is due to differences in their methionine contents. We have also compared the protein sequence of the Brazil nut sulfur-rich protein to that of the 15 kDa high sulfur zein protein from maize [18] which contains about 11% methionine and have found no significant homology between these two proteins.

Discussion

The majority of known proteins, of both plant and animal origin, have relatively low levels of methionine, usually around 1–2% as predicted by the theory of molecular evolution [16]. In the present study, we have partially sequenced an abundant protein from Brazil nuts which is exceptionally rich in methionine (18%) and have identified and sequenced a cDNA clone encoding this protein. Only one plant protein with comparable levels of methionine has been reported in the literature. Phillips and McClure recently described the amino acid composition of a polypeptide of 10 kDa in a maize mutant, BSSS-53, containing 21 mol% methionine [19].

The sequence data from the Brazil nut cDNA clone as well as the data from the hybrid-selected translation experiment are consistent with previous

in vitro translation studies which have shown that the sulfur-rich protein is synthesized as a larger precursor polypeptide. The size of the polypeptide encoded by pHS-3 would be about 17 kDa, which is close to the 18 kDa value for the precursor obtained from the sizing on polyacrylamide gels of the polypeptides translated from Brazil nut RNA *in vitro* [2]. The correlation of the amino acid sequence obtained from the purified 9 kDa subunit with the last 77 amino acids of the sequence derived from the nucleotide sequence indicates that the processing steps which are involved in the maturation of the sulfur-rich protein must be taking place at the amino terminal end of the precursor. Previous *in vivo* labelling studies demonstrated that there are 3 distinct processing steps. First, a small peptide, most likely a signal sequence, is cleaved from the 18 kDa precursor to generate a 15 kDa polypeptide which subsequently is processed to a 12 kDa polypeptide and then to the 9 kDa and 3 kDa subunits [2]. We have not determined experimentally the precise residues which are cleaved upon maturation of the sulfur-rich protein. Nonetheless, we can propose approximate cleavage sites that would divide the amino acid sequence into four domains corresponding to the observed polypeptides (Fig. 2A). The hydrophobic nature of the amino terminus of the encoded polypeptide (Fig. 2B) suggests that this region serves as a signal peptide. The alanine and phenylalanine residues at positions #22 and 23 would represent a possible cleavage site for a signal peptidase as determined by the $(-3-1)$ rule of Von Heijne [25]. A second cleavage may take place around amino acid residue #46 and would result in a polypeptide of about 12 kDa. We have attempted to determine the exact location of this cleavage site by sequencing the 12 kDa precursor polypeptide, but found that its NH_2 terminus is blocked. Finally, the major (80%) partial amino acid sequence of the 9 kDa subunit would predict that the cleavage site for the third processing step is between methionine residue #69 and proline residue #70, whereas the minor sequence (20%) would indicate that the final processing site is three amino acids away, between residues #72 and 73. The 3 kDa region clipped off in this final processing step is extremely rich in methionine

(~20%). We suspect that this portion of the precursor accumulates and gives rise to the 3 kDa subunit of the sulfur-rich protein. *In vivo* labelling studies using [^{35}S]methionine indicate that the 3 kDa subunit is rich in methionine (data not shown). In addition, the amino acid composition of the sulfur-rich protein supports this notion. Tyrosine and threonine residues are present in the amino acid analysis of the purified sulfur-rich protein (9 kDa + 3 kDa) (data not shown). These residues are not found in the amino acid sequence derived from the nucleotide sequence of the 9 kDa subunit but are present in the 3 kDa region immediately preceding the 9 kDa subunit.

The homology between the sulfur-rich protein from Brazil nut and seed proteins from castor bean and rapeseed is particularly striking since the three plants are not closely related taxonomically and the castor bean and rapeseed proteins contain low levels of methionine. The proteins from all three plants consist of a small and a large subunit polypeptide and contain high levels of cysteine. The positions of these cysteine residues are conserved, suggesting that the structural frameworks of these three proteins may be quite similar despite the dramatic differences in their methionine contents. This structural similarity may be conserved in the small water-soluble proteins in other oilseeds of diverse phylogenetic relationships as well. In a survey of the amino acid compositions of 2S seed proteins, Youle and Huang [26] noted that the levels of cysteine in proteins from different oilseeds (sunflower, mustard, linseed, lupin, cucumber, Brazil nut, hazelnut, yucca, castor bean, and cotton) were quite high and in fact very similar. Because of their high amide contents, abundance in seeds, and disappearance from seeds during germination, these low molecular weight proteins were suggested to function as seed storage proteins with the additional and unique role of providing sulfur reserves for germination [26]. Of these proteins, however, only the Brazil nut 2S protein contains unusually high levels of methionine, contrary to theoretical predictions based on the theory of molecular evolution [16]. At the present time, we do not know why Brazil nuts might require such high levels of methionine. The soil in the Amazon region is rather poor

in sulfur [21]; possibly these levels of methionine are required in order to provide an adequate supply of methionine to the germinating seeds. Whatever the function of the Brazil nut sulfur-rich protein, it appears that the structural framework of the 2S seed proteins is flexible enough to accommodate large numbers of methionine residues while still preserving the small size, water solubility, and high amide content of these proteins.

Both the castor bean protein and the rapeseed protein are analogous to the Brazil nut sulfur-rich protein in that they are composed of two low molecular weight subunits. In the case of castor bean, the large subunit of the protein is homologous with the 9 kDa subunit portion of the Brazil nut precursor polypeptide while the small subunit of the castor bean protein appears to correspond to the region of the Brazil nut protein which we believe encodes the 3 kDa subunit. Interestingly, the junction between the large and small subunits of the castor bean protein corresponds to the minor cleavage site of the Brazil nut 12 kDa precursor (amino acid residue #72) (Fig. 4B). These data suggest that both subunits of the castor bean protein

may be synthesized as part of a larger precursor similar to the Brazil nut sulfur-rich protein and that the final processing step involved in the maturation of the castor bean protein may be similar to that found with the Brazil nut protein.

The processing involved in the maturation of the rapeseed protein [5] also bears similarities to that of the Brazil nut sulfur-rich protein. As with the Brazil nut protein, the large subunit of napin is found at the carboxyl terminal portion of the precursor (Fig. 4B). In both Brazil nut and rapeseed, the precursor polypeptide undergoes extensive processing before reaching its mature subunits. From the best alignment of the amino acid sequence of the large subunit of the rapeseed protein with that of the Brazil nut protein sequence, it appears that the cleavage site of the large subunit of napin occurs at about the same point as the primary cleavage site of the Brazil nut large subunit (amino acid residue #69) (Fig. 4B).

In the past, there has been much effort to enhance the sulfur amino acid content of seeds, particularly those from legumes, by conventional plant breeding approaches. The overall improvement in

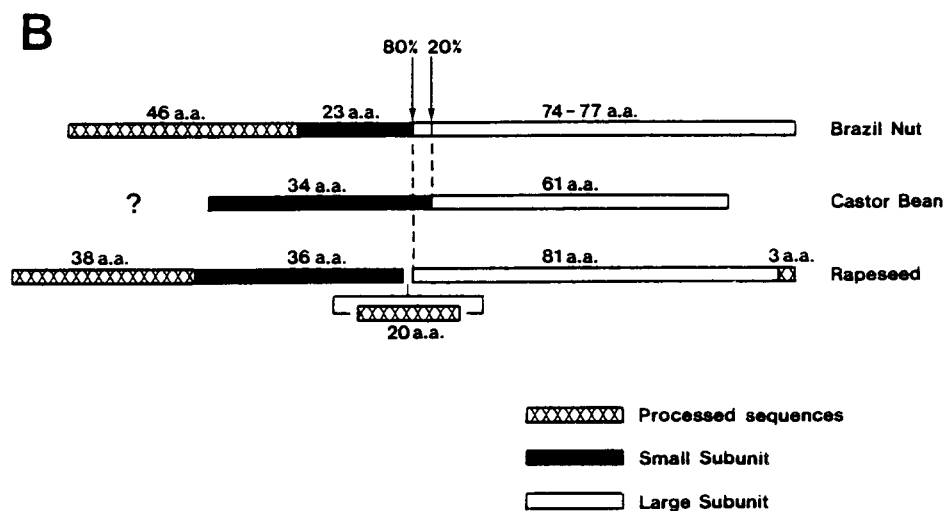


Fig. 4B. Comparison of the processing schemes utilized in the maturation of the subunits of the 2S water-soluble seed proteins of Brazil nut, castor bean and rapeseed showing the positions of the large and small subunits of the castor bean and rapeseed proteins relative to the Brazil nut sulfur-rich protein. The locations of the final processing sites involved in the maturation of the Brazil nut sulfur-rich protein are indicated with arrows. The large subunit of the castor bean protein begins at the same amino acid residue as 20% of the large subunit molecules from the Brazil nut protein, whereas the processing site in the rapeseed protein appears to correlate with the 80% processing site from Brazil nut.

the nutritional quality of these seeds has not been significant [17], although the same approach was successful in obtaining high lysine corn [14, 15]. Studies of seed proteins in oilseeds have shown that there is a wide occurrence of abundant 2S proteins in diverse plant species. These proteins appear to have similarity in their structural framework and precursor processing, seem to serve a storage function, and have a seemingly flexible amino acid composition. The fact that a large amount of methionine is localized in a single 2S protein species in Brazil nut suggests to us a molecular approach for improving the nutritional quality of seed proteins deficient in the sulfur amino acids. The cloning of a cDNA encoding this sulfur-rich protein thus represents a first step in an effort to alter the amino acid composition of seed proteins. A further understanding of the genes which encode this unusual sulfur-rich protein should provide additional useful information.

Acknowledgements

We thank Mr. Bruce Nelson for help in collecting the Brazil nut fruits, Mr. Alan Smith (Department of Biochemistry, University of California, Davis, CA) for protein sequence determination, Ms. Kathy Mead and Dr Danny Alexander for the synthesis of the oligodeoxynucleotide used in the primer extension analysis, and Dr Phil Filner for helpful suggestions and critical review of this manuscript.

References

- Alexander DC, McKnight TD, Williams BG: A simplified and efficient vector-primer cDNA cloning system: *Gene* 31:79–89, 1984.
- Altenbach SB, Pearson KW, Leung FW, Sun SSM: The step-wise processing of a sulfur-rich seed protein from Brazil nut (*Bertholletia excelsa*). In: Shannon LM, Chrispeels MJ (eds) *Molecular Biology of Seed Storage Proteins and Lectins*. Waverly Press, Baltimore, Maryland, 1986, pp 137–146.
- Antunes AJ, Markakis PJ: Protein supplementation of navy beans with Brazil nuts. *Agric Food Chem* 25:1096–1098, 1977.
- Bhown AS, Mole JE, Weissinger A, Bennett JC: Methanol solvent system for rapid analysis of phenylthiohydantoin amino-acids by high pressure liquid chromatography. *J Chromatogr* 148:532–535, 1978.
- Crouch ML, Tenbarger KM, Simon AE, Ferl R: cDNA clones for *Brassica napus* seed storage proteins: evidence from nucleotide sequence analysis that both subunits of napin are cleaved from a precursor polypeptide. *J Mol & Appl Genetics* 2:273–283, 1983.
- Edman P, Begg G: A protein sequenator. *Eur J Biochem* 1:80–90, 1967.
- Hall TC, Ma Y, Buchbinder BU, Pyne JW, Sun SSM, Bliss FA: Messenger RNA for G1 protein of French bean seeds: cell-free translation and product characterization. *Proc Natl Acad Sci USA* 75:3196–3200, 1978.
- Hopp TP, Woods KR: Prediction of protein antigenic determinants from amino acid sequences. *Proc Natl Acad Sci USA* 78:3824–3828, 1981.
- Kulbe DK: Micro polyamide thin layer chromatography of phenylthiohydantoin amino-acids at subnanomolar level, a rapid microtechnique for simultaneous multi-sample identification after automated Edman degradations. *Anal Biochem* 59:564–573, 1974.
- Kozak M: Comparison of initiation of protein synthesis in prokaryotes, eukaryotes, and organelles. *Microbiol Rev* 47:1–45, 1983.
- Laemmli UK: Cleavage of structural proteins during the assembly of the head of bacteriophage T-4. *Nature* 227:680–685, 1970.
- Maniatis T, Fritsch EF, Sambrook J: *Molecular Cloning. A Laboratory Manual*. Cold Spring Harbor Laboratory, New York, 1982, pp 331–333.
- Maxam AM, Gilbert W: A new method for sequencing DNA. *Proc Natl Acad Sci USA* 74:560–564, 1977.
- Mertz ET, Bates LS, Nelson OE: Mutant gene that changes protein composition and increases lysine content of maize endosperm. *Science* 145:279–280, 1964.
- Nelson OE, Mertz ET, Bates LS: Second mutant gene affecting the amino acid pattern of maize endosperm proteins. *Science* 150:1469–1470, 1965.
- Ohta T, Kimura M: Amino-acid composition of proteins as a product of molecular evolution. *Science* 174:150–153, 1971.
- Payne PI: Breeding for protein quantity and protein quality in seed crops. In: Daussant J, Mosse J, Vaughan J (eds) *Seed Proteins*. Academic Press Inc., London, 1983, pp 223–253.
- Pedersen K, Argos P, Naravana SVL, Larkins B: Sequence analysis and characterization of a maize gene encoding a high-sulfur zein protein of M_r 15000. *J Biol Chem* 261:6279–6284, 1986.
- Phillips L, McClure BA: Elevated protein-bound methionine in seeds of a maize *Zea mays* line resistant to lysine plus threonine. *Cereal Chem* 62:213–218, 1985.
- Pisano JJ, Bronzert TJ, Brewer HB Jr: Advances in the gas chromatographic analysis of amino-acid phenylthiohydantoin and methylthiohydantoin. *Anal Biochem* 45:43–59, 1972.

21. Sanchez PA, Bandy DE, Villachica JH, Nicholaides JJ: Amazon basin: management for continuous crop production. *Science* 216:821–827, 1982.
22. Sanger F, Nicklen S, Coulson AR: DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci USA* 74:5463–5467, 1977.
23. Sharief FS, Li SS-L: Amino acid sequence of small and large subunits of seed storage protein from *Ricinus communis*. *J Biol Chem* 257:14753–14759, 1982.
24. Taub F, Thompson EB: An improved method for preparing large arrays of bacterial colonies containing plasmids for hybridization: *in situ* purification and stable bonding of DNA on paper filters. *Anal Biochem* 126:222–230, 1982.
25. Von Heijne G: How signal sequences maintain cleavage specificity. *J Mol Biol* 173:243–251, 1984.
26. Youle RJ, Huang AH: Occurrence of low molecular weight and high cysteine-containing albumin storage proteins in oilseeds of diverse species. *Amer J Bot* 68:44–48, 1981.