Isolation and Characterization of 2S Cocoa Seed Albumin Storage Polypeptide and the Corresponding cDNA

Sunil Kochhar,*,† Karin Gartenmann,† M. Guilloteau,§ and J. McCarthy§

Nestlé Research Center, P.O. Box 44, Vers-chez-les-Blanc, CH-1000 Lausanne 26, Switzerland, and Nestlé Research Center Tours, Avenue Gustave Eiffel, Notre Dame d'Oé, B.P. 9716, 37097 Tours Cedex 2, France

The amine pool of cocoa is known to be an essential component for the development of the typical cocoa flavor. To better understand and to produce an intense in vitro cocoa flavor, identification of the polypeptides that are the source of the amine flavor precursor pool is essential. Chromatographic analysis of the polypeptide profile of unfermented cocoa resulted in identification of a novel storage polypeptide of M_r 8515. The N-terminal sequence of the first 34 residues of the purified polypeptide shows similarity to 2S storage albumins of cotton and Brazil nut and sweet protein, Mabinlin. To identify the corresponding cDNA of the putative cocoa 2S albumin, 18 randomly chosen clones from the cDNA library of immature *Theobroma cacao* seed mRNA were sequenced, and a full-length cDNA clone encoding a protein harboring the N-terminal sequence of the novel polypeptide was selected. The open reading frame of the clone encodes a polypeptide of M_r 17125. Comparison of the translated amino acid sequence of the precursor protein or the mature polypeptide against the Swiss-Prot and TrEMBL databases shows high sequence similarity (>52%) and identity (>38%) to many plant 2S albumins. Tryptic peptide mass fingerprinting of the purified polypeptide by high-performance liquid chromatography−electrospray ionization mass spectrometry shows 10 masses that match the expected tryptic peptides of the deduced sequence. Together with the published work on plant 2S albumin processing, the results presented here suggest that post-translational processing yields a 73-residue polypeptide (residue positions 78−150) corresponding to the 9 kDa subunit of the mature cocoa 2S albumin protein.

Keywords: *2S albumin; albumin; LC-MS; tryptic peptides; processing sites; storage proteins; Theobroma cacao*

INTRODUCTION

An important part of the specific flavor of cocoa arises when peptides and free amino acids present in fermented cocoa beans undergo complex Maillard reactions with the sugar molecules also present in the beans (*1*−*7*). The majority of these Maillard reactions occur during the drying and roasting stages of bean processing. There are few cocoa flavor precursor peptides and free amino acids present in the fresh cocoa bean. However, a significant amount of these precursors is present after fermentation of the beans (*8*, *9*). It is now well established that these cocoa flavor precursors are produced during fermentation as a result of acid-induced extensive proteolysis of cocoa bean proteins (*4*, *5*). Although it is known that hydrophobic amino acids are important cocoa flavor precursors (*9*, *10*), the specific peptides responsible for generating cocoa flavor during roasting remain uncharacterized. To better understand cocoa and chocolate flavor, it is necessary to identify and characterize the flavor precursors in the cocoa seed.

Proteins constitute 10−15% dry weight of cocoa seeds, the second most abundant constituent after cocoa fat. Voigt et al. (*3*), employing sodium dodecyl sulfate−polyacrylamide gel electrophoresis (SDS-PAGE), showed that four predominant proteins represent 95% (w/w) of the total cocoa seed protein. Furthermore, it was suggested that the total seed protein content is composed of 52 and 43% of albumin and globulin fractions, respectively (*3*). Later, on the basis of analysis of silver-stained 2D SDS-PAGE gels, it was estimated that the vicilin protein (31 and 47 kDa polypeptides) could constitute >30% and the albumin (21 kDa polypeptide) another 15−30% in the mature cocoa seed protein (*11*). In addition, electrophoretic patterns of total cocoa proteins showed that there are a number of other relatively highly expressed proteins in cocoa seed (*11*). It is quite possible that one or more of these other relatively abundant cocoa seed proteins also contain cocoa flavor precursor peptides.

In 1985, degradation of the vicilin storage polypeptides (31 and 47 kDa) was followed during germination, and it has been suggested that the specific degradation of these two abundant polypeptides during fermentation may generate flavor precursor peptides (*9*). Subsequently, by cloning a cDNA clone encoding both polypeptides, it was shown that the cocoa vicilin protein was the precursor of these two seed storage polypeptides (*12*, *13*). More recently, it has been shown that when the vicilin storage polypeptides are treated with partially purified cocoa seed aspartic proteinase and cocoa seed carboxypeptidase, precursors of a cocoa flavor can be generated (*4*, *5*).

In this paper the isolation and characterization of a cocoa 2S seed storage polypeptide and its corresponding cDNA clone are presented. During fermentation, the

* Corresponding author (telephone +41 21 785 9336; fax +41 21 785 8549; e-mail sunil.kochhar@rdls.nestle.com).
† Nestlé Research Center Lausanne.
§ Nestlé Research Center Tours.

10.1021/jf010497b CCC: $20.00
Published on Web 08/22/2001

cocoa seed proteins, including the 2S albumin, undergo extensive proteolysis and thus contribute to the total cocoa amine flavor precursor pool.

EXPERIMENTAL PROCEDURES

Materials. HPLC grade acetonitrile, methanol, and water were from Merck. Ethylenediaminetetraacetic acid sodium salt (EDTA), tripotassium phosphate monohydrate, 1,4-dithio-DL-threitol (DTT), trifluoroacetic acid (TFA), and guanidinium hydrochloride (Gnd-HCl) were from Fluka. 4-Vinylpyridine was from Aldrich. All other chemicals used were of analytical grade. Unless stated otherwise, all studies were carried out using West African Amelonado cocoa beans. For the isolation of cocoa seed mRNA, an immature green pod of EET-95 (with yellow spots) was obtained from the greenhouse at Nestlé Research Center Tours, Tours, France. This pod was the result of open pollination. Cocoa acetone powder (CAP) was prepared from non-defatted cocoa beans according to the method of Voigt et al. (*3*). For details see ref *14*.

Preparation of mRNA from Cocoa Seeds. The seeds were extracted from the immature pod of EET-95. The seeds in the pod displayed two very distinct developmental stages. One set of seeds was more mature; that is, the seeds were purple with only small amounts of white gelatinous matrix tissue in seed folds. The other set of seeds was partially light pink and partially white and had significant amounts of the gelatinous matrix material in the folds of the developing seed. For RNA isolation, small pieces from three of the more mature seeds and small pieces from two less mature seeds were taken immediately after the seeds were freed from the matrix, and these seed parts were directly frozen in liquid nitrogen. The frozen seed material was then ground to a powder with a pestle in the presence of liquid nitrogen. The liquid nitrogen plus cocoa powder was put in a 50 mL Falcon tube, and the liquid nitrogen was allowed to evaporate. As the powder warmed toward 0 °C, 28 mL of solution A was added [14 mL of 100 mM Tris-HCl, pH 8; 14 mL of Aqua Phenol (Appligene-Oncor); 0.1% hydroxyquinoline; 140 μL of 10% SDS; 110 μL of β-mercaptoethanol]. This mixture was homogenized with a glass dunce homogenizer on ice. The resulting solution was spun for 10 min at 6720*g*. The aqueous phase was recovered and manually mixed with 7 mL of phenol plus 7 mL of chloroform/isoamyl alcohol (Ready Red; Appligene-Oncor). The extraction was then spun at 6720*g* for 10 min. After this stage, great care was taken to avoid any RNase contamination of the sample. The aqueous phase recovered was re-extracted twice more with 14 mL of Ready Red. The final aqueous phase obtained was made 0.3 M in sodium acetate, and 2 volumes of 100% ethanol was added. The tube was then mixed and put at −20 °C for 1 h, at −80 °C for 15 min, and then spun for 30 min at 6720*g*.

The nucleic acid pellet recovered was slowly resuspended in 10 mL of 100 mM Tris-HCl, pH 8. Then, 3 mL of 8 M lithium chloride was added (1.625 M final) followed by 2 volumes of ethanol. This mixture was put for 1 h at −20 °C followed by 15 min at −80 °C. The nucleic acid precipitate was recovered by centrifugation at 6720*g* for 30 min. This pellet was resuspended in 600 μL of RNase-free water, aliquoted into small samples, and frozen at −80 °C.

Preparation of cDNA Library from Cocoa Seed mRNA. Poly A+ RNA was prepared from the total cocoa seed RNA prepared as described above using the Oligotex kit from Qiagen. The procedure employed was exactly as described in the kit instructions for 250−500 μg total RNA. The synthesis of cDNA from the poly A+ mRNA was carried out using a SMART PCR cDNA synthesis kit from Clontech according to the kit instructions using Gibco BRL Superscript II MMLV reverse transcriptase. The cDNA obtained was then blunt end ligated into the PCR-Script Amp SK(+) cloning vector of Stratagene. The ligated DNA was transformed into Stratagene Ultracompetent cells XL-2 Blue as described in the instruction manual for these cells.

Preparation of Gnd-HCl Extract. CAP (300 mg) was suspended in 3 mL of extraction buffer (100 mM ammonium phosphate, 66.7 mM potassium hydroxide, 3 mM EDTA, and 6 M Gnd-HCl) and sonicated for 1 min (10 mm probe at medium setting and full power, Labsonic U, B. Braun). The suspension was cooled on ice for 15−30 min and centrifuged at 15000*g* at 4 °C for 15 min. *S*-Pyridinylethylation of the CAP extract was carried by sparging the CAP extract (2 mL) with argon and mixing with 10 μL of reducing solution (0.8 M DTT in 3 M tripotassium phosphate/3 mM EDTA). The solution was kept at room temperature in the dark for 60 min. Ten microliters of 4-vinylpyridine was added and the solution mixed vigorously. The reaction mixture was further incubated for 30 min at room temperature, passed through a 0.22 μm filter disk, and kept at 4 °C until analyzed (see ref *15*).

Protein Assay. Protein concentrations in the cocoa extract were determined according to the method of Lowry et al. (*16*) as modified by Bio-Rad (DC protein assay kit). Bovine serum albumin was used as a protein standard.

Purification of Cocoa Polypeptides by RP-HPLC. All chromatographic separations were carried out at room temperature using the Bio-Rad HRLC series 800 system consisting of two high-pressure pumps (model 1350 series), an autosampler (model AS 100-T) equipped with a Rheodyne injection valve (sample loop, 1 mL), a dynamic gradient mixer, and a UV−vis detector (model 1090). ValueChrom chromatography software (Bio-Rad) controlled the HPLC system. Solvents were degassed using an on-line degasser (Gastorr 102). Reduced and pyridinylethylated Gnd-HCl extracts of CAP were injected onto a reversed-phase C_4 HPLC column [Vydac protein C4 (5 μm, 4.6 × 220 mm)] pre-equilibrated with solvent A [0.1% TFA (v/v) in water] and eluted with a linear gradient of increasing concentration of solvent B [0.1% TFA/80% ACN (v/v) in water]: 0−15% B in 5 min, 15−27% B in 40 min, 27−35% B in 2 min, isocratic at 35% B for 3 min, 35−43% B in 25 min, 43−56% in 50 min, 56−70% in 5 min, 70−100% in 10 min, and finally isocratic at 100% B for 5 min. The flow rate was 1 mL/min, and fractions of 1 mL each were collected automatically (Frac-100, Pharmacia). Detection of peaks was carried out at 215 nm. For preparative purification of the 9 kDa polypeptide, a semipreparative C_4 reversed-phase HPLC column [Vydac protein C4 (10 μm, 22 × 250 mm)] was employed with the same solvent system and gradient elution except the flow rate was 4 mL/min.

PAGE Analysis. Denaturating SDS-PAGE was carried out using gradient ready gels [Bio-Rad, 10 × 10 cm, Tris-HCl 8−16% T (*17*) and Tris-Tricine 10−20% (*18*)] using the Miniprotean 3 system from Bio-Rad. Glycosylation of the purified polypeptide was assessed by employing the glycoprotein detection kit from Bio-Rad.

Generation of Tryptic Peptides. The purified reduced and pyridinylethylated polypeptide was dried under reduced pressure and dissolved in 100 μL of 50 mM Tris-HCl buffer, pH 8.5, containing 2 mM $CaCl_2$. Digestions with trypsin (TPCK treated, Promega) were carried out at 25 °C overnight at an enzyme substrate ratio of 1:20 and stopped by the addition of TFA to a final concentration of 1% (v/v).

Amino Acid Analysis. Purified polypeptide was acid hydrolyzed in 6 N HCl for 22 h at 110 °C under argon. The amino acid analysis was carried out by RP-HPLC following precolumn derivatization with phenylisothiocyanate (PITC) according to the method of Heinrikson and Meredith (*19*).

Amino Acid Sequencing. Purified cocoa polypeptide was subjected to N-terminal amino acid sequencing by Edman degradation employing a gas phase sequencer (Procise 494, Perkin-Elmer) using standard cycles and standard methods.

HPLC-ESI-MS Analysis. MS and MS-MS measurements were made using a Finnigan-MAT LCQ mass spectrometer interfaced with a Spectra HPLC system (Finnigan-MAT) described in detail elsewhere (*14*, *20*).

The LC-MS analysis of the intact and pyridinylethylated albumin was carried out using an RP C_8 column [Spherisorb 80-5 C_8 (5 μm, 2 × 125 mm), Macherey-Nagel] with a linear increase of solvent B [0.05% TFA/80% ACN (v/v) in water] in solvent A [0.045% TFA (v/v) in water] as follows: 0−35% B in 5 min, isocratic at 35% B for 5 min, 35−60% B in 70 min, 60−100% B in 10 min, and finally isocratic at 100% B for 5 min.

```
AAGCAGTGGTAACAACGCAGAGTACGCGGGGAAGAACCAAAGCCTTGTCA 50
TCTAACTAGCTATATATCTATATCCACCATGGCAAAGCTCGGTCTCCTCC 100
                          M  A  K  L  G  L  L  L

TAGCCACCCTTGCTCTTGTTCTCTTCCTCGGCAATGCCTCCGTTTACCAC 150
  A  T  L  A  L  V  L  F  L  G  N  A  S  V  Y  H

ACCACCGTCACGGTTGACAGCGAGGAAAACCCTTGGGGAAGCAAAGAGAG 200
T  T  V  T  V  D  S  E  E  N  P  W  G  S  K  E  S

CAGCTGTCAGAAGCAGATAAAGAAGCAAAACTACCTCAAGCACTGTCAGG 250
 S  C  Q  K  Q  I  K  K  Q  N  Y  L  K  H  C  Q  E
AGTACATGGAGGAGCAGTCCAGAGGCAGCGGCAGCAGCAGCCGTGAG 300
  Y  M  E  E  Q  S  R  G  S  G  S  S  S  S  R  E

CGCTACAGCCGCCCCGTGAGCAAGCACCTAGACTCCTGTTGCCAGCAACT 350
R  Y  S  R  P  V  S  K  H  L  D  S  C  C  Q  Q  L

GGAGAAGCTCGATACGCCGTGCCGTTGCCCTGGTCTAAAACAGGCAGTGC 400
 E  K  L  D  T  P  C  R  C  P  G  L  K  Q  A  V  Q

AGCAACAGGCGGAAGAGGGAGAGTTTGGGAGGGAAGAGTTGCAAGAGATG 450
  Q  Q  A  E  E  G  E  F  G  R  E  E  L  Q  E  M

TATGAGACGGTTGACAAGATCATGAACAAGTGTGACGTAGAGCCTGGAAG 500
Y  E  T  V  D  K  I  M  N  K  C  D  V  E  P  G  R

GTGTAACTTGCAACCTCGCAACTGGTTCTAGAGAGAAAGAAAGATCAGAG 550
 C  N  L  Q  P  R  N  W  F  *

CTGCCTGATCTAATGTAAAACAATGACTGTAATGTTTCACCCATCAACTC 600
TGGTGTTCTAACTGGAGGTTTTTGGGGTGACTAGAAGTACTGATAATCCA 650
TAAATAAAAGCACATTCTCGTGTGCGGTTGCTTTTTGCTTCAGGCCAAAA 700
AAAAAAAAAAAAAAAAA                                   718
```

Figure 1. DNA sequence of the cDNA clone encoding the putative 2S albumin precursor. The underlined sequence corresponds to the N-terminal sequence of the 9 kDa polypeptide.

The flow rate was 0.2 mL/min and UV detection at 215 nm. Tryptic peptides of reduced and pyridinylethylated polypeptides were analyzed using an RP C_{18} column [Nucleosil 100-3 C18 HD (3 μm, 2 $\times$ 150 mm), Macherey-Nagel] with a linear increase of solvent B in solvent A as follows: 0−60% B in 60 min, 60−100% B in 20 min, followed by an isocratic elution at 100% B for 5 min. The flow rate was 0.2 mL/min and detection at 215 nm.

Database Analysis. The theoretical tryptic mass fingerprint analysis and the comparisons of peptide masses against Swiss-Prot and TrEMBL databases were carried out using software tools PeptideMass and PeptIdent available at ExPASy (www.expasy.ch) at the Swiss Institute of Bioinformatics, Geneva, Switzerland. The deduced amino acid sequence of cocoa 2S albumin was compared with the protein databases Swiss-Prot and TrEMBL using the GCG program FASTA [Genetics Computer Groups (GCG), Madison, WI]. Pairwise comparisons were carried out using GCG programs GAP and PILEUP.

RESULTS AND DISCUSSION

Identification and Analysis of cDNA Encoding Cocoa 2S Albumin Precursor. To identify and characterize the DNA sequences encoding storage proteins, a cDNA library was constructed from the mRNA isolated from immature cocoa seeds. Sequencing of randomly chosen clones using the T3 primer present in the vector allowed the identification of a cDNA clone that encodes an amino acid sequence with significant homology to other plant 2S albumin sequences. This clone was then sequenced entirely on both strands. The DNA sequence, and the translated amino acid sequence, is shown in Figure 1. The DNA insert is 718 base pairs, and analysis of the protein encoded by this cDNA shows a calculated molecular weight of 17125 Da and a p*I* of 6.15.

A computer search of the Swiss-Prot and TrEMBL databases using software FASTA (*21*, *22*) and the deduced amino acid sequence of the complete cDNA clone as the query sequence shows similarity to mainly plant 2S albumins. Rigorous pairwise sequence comparisons by GAP algorithm with an optimized gap penalty of 10 reveals >30% amino acid identity over many storage 2S albumins. The cotton 2S albumin is most similar, showing 55% identity, and when the conservative mutations of the amino acid residues with similar properties, that is, size, charge, or hydrophobicity, are taken into consideration, both sequences are 68% similar. In addition, the cocoa polypeptide is homologous to 2S albumins from English walnut, Brazil nut, Arabdopsis, pumpkin, rape seed, and sweet protein Mabinlin II. The amino acid identity within the maximum overlap is >30% (Table 1). The observation that the cocoa 2S cDNA amino acid sequence is most homologous to the cotton 2S albumin supports an earlier proposal that cocoa is very closely related to cotton because the cocoa vicilin storage protein is most homologous with the cotton vicilin sequence (*13*).

The aligned sequences of the similar 2S plant albumins are presented in Figure 2. The most striking conserved feature observed between the 2S sequences is the position and number of cysteine residues. All of the sequences have eight cysteines, and these cysteines exhibit the common arrangement of ...C..C...CC...CXC...C...C previously identified by Rico et al. (*23*). Although the cotton sequence has a number of regions of sequence identity of three to six amino acids with the cocoa protein sequence (plus one region of nine amino acids), in general, the 2S sequences examined exhibit relatively few short contiguous regions of sequence identity with the cocoa sequence. This observation indicates that if the cocoa polypeptide contains one or more flavor peptide sequences, these peptide sequences are probably unique to the cocoa 2S protein.

Purification and Characterization of Major Cocoa Polypeptides. RP-HPLC of the reduced and pyridinylethylated 6 M Gnd-HCl extract of the cocoa acetone powder shows a total of five major polypeptide peaks (Figure 3). SDS-PAGE analysis of the pooled fractions reveals that the polypeptide peak eluting at a retention time of 78 min (pool C) is the abundant 21 kDa albumin, and those eluting at 87 and 91 min (pools D and E) are the vicilin storage polypeptides (31 and 47 kDa). The polypeptides eluting at retention times of 37 min (pool A) and 68 min (pool B) focus as 12 and 9 kDa proteins in Tricine-SDS-PAGE. Both of the polypeptides, to the best of our knowledge, are unknown cocoa storage polypeptides. The 9 kDa polypeptide was purified from the Gnd-HCl extract with or without *S*-pyridinylethylation employing a semipreparative C_4 reversed-phase column. The elution profile of cocoa polypeptides of the reduced and pyridinylethylated extract was similar to those obtained from the underivatized extract except that the polypeptides eluted 3−5 min earlier (data not shown). The 9 kDa polypeptide (pool B) was further purified using a C_4 reversed-phase column (Figure 3B). Analysis of the purified polypeptide by Tricine-SDS-PAGE shows a single protein band (Figure 4), indicating homogeneity of the preparation.

LC-ESI-MS analysis shows the molecular weight of the mature protein is 8513 $\pm$ 2 Da (Figure 5). Reduction and *S*-pyridinylethylation result in a positive shift of 630 mass units (M_r 9143), indicating the presence of six cysteine residues (Figure 5). The purified polypeptide,

Table 1. Percentage Identity and Similarity among Different Plant 2S Albumins[a]

	*Tca*_2S	*Ghi*_2S	*Jre*_2S	*Bex*_2S	*Ath*_2S	*Cam*_2S	*Cma*_2S	*Bra*_2S
*Tca*_2S		**55**	**38**	**36**	**36**	**34**	**31**	**36**
*Ghi*_2S	68		**39**	**39**	**38**	**37**	**37**	**33**
*Jre*_2S	53	51		**44**	**33**	**39**	**46**	**35**
*Bex*_2S	47	49	54		**43**	**32**	**37**	**27**
*Ath*_2S	50	47	44	50		**49**	**35**	**68**
*Cam*_2S	45	46	48	42	63		**36**	**48**
*Cma*_2S	46	49	59	53	48	49		**48**
*Bra*_2S	50	43	49	40	74	64	48	

[a] Percent identity, i.e., invariant residues (in bold letters) and similarity, i.e., conservative mutations were calculated by alignment of two sequences employing software GAP. The gap weight values were set to 8. The abbreviations of the names of the proteins are given in Figure 5.

```
        1                                                  50
Ara_2S  MANKLFLVCA TLALCFLLTN ASIYRTVVEF EEDDASNPVG PRQ..RCQKE
Bra_2S  MANKLFLVSA TLAFFFLLTN ASIYRTIVEV DEDDATNPAG PFRIPKCRKE
Mab_2S  MAKLIFLF.A TLALFVLLAN ASIQTTVIEV DEEEDN.... ..QLWRCQRQ
Tca_2S  ~MAKLGLLLA TLALVLFLGN ASVYHTTVT. ..VDSEENPW GSKESSCQKQ
Ghi_2S  ~MAKLAVYLA TLALILFLAN ASI..TSVT. ..VESEEN.. ...RDSCEQQ
Ber_2S  ~MAKISVAAA ALLVLMALGH ATAFRATVTT TVVEEE.... ..NQEECREQ
Jre_2S  ~~~~~~~~AA LLVALLFVAN AAAFRTTITT MEIDEDIDNP RRRGEGCREQ
Cma_2S  ~MARLTSIIA LFAVALLVAD AYAYRTTITT VEVEE...NR QGREERCR.Q

       51                                                 100
Ara_2S FQQSQHLRAC QRWMSKQMRQ GRGGGPS..L DDEFDFE... ....GPQQGY
Bra_2S FQQAQHLKAC QQWLHKQAMQ S.GSGPSWTL DGEFDFEDDM ENPQGPQQRP
Mab_2S FLQHQRLRAC QRFIHRRAQ. .FGGQPD.EL EDEVEDDNDD EN....QPRR
Tca_2S IKKQNYLKHC QEYMEEQSRG SG........ .......SSS SRERYSRPVS
Ghi_2S IRKQAHLKHC QKYMEEEWGG EG........ .......SDN IAGGY.....
Ber_2S MQRQQMLSHC RMYMRQQMEE SP........ .......YQ. ..TMPRRGME
Jre_2S IQRQQNLNHC QYYLRQQSRS GG........ .......YD. ..EDNQR...
Cma_2S MSAREELRSC EQYLRQQSRD VL........ .......QMR GIENPWRREG

       101                                                150
Ara_2S QLLQQCCNEL RQEEPVCVCP TLKQAARAV. ....SLQGQH G.PFQS..RK
Bra_2S PLLQQCCNEL HQEEPLCVCP TLKGASKAVK QQVRQQQGQQ GQQLQQVISR
Mab_2S PALRQCCNQL RQVDRPCVCP VLRQAA.... QQVLQRQIIQ GP...QQLRR
Tca_2S KHLDSCCQQL EKLDTPCRCP GLKQAV.... ..QQQAE..E GEFGREELQE
Ghi_2S ..IDSCCQQL EKMDTQCRCQ GLRHAT.... ..MQQMQQMQ EQMGSKQMRE
Ber_2S PHMSECCEQL EGMDESCRCE GLRMMM.... ..MRMQQEEM QPRG.EQMRR
Jre_2S QHFRQCCQQL SQMDEQCQCE GLR.QV.... ..VRRQQQQQ GLRG.EEMEE
Cma_2S GSFDECCREL KNVDEECRCD MLEEIA.... .....REEQR QARG.QEGRQ

       151                                 184
Ara_2S IYQSA.KYLP NICKIQQVGE CPFQTTIPFF PPYY
Bra_2S IYQTA.THLP KVCNIPQVSV CPFQKTMP.G PSY~
Mab_2S LFDAA.RNLP NICNIPNIGT CPFR.TWP~~ ~~~~
Tca_2S MYETVDK.IM NKCDV.EPGR CNLQPRNWF~ ~~~~
Ghi_2S IMQKVTKNIM SECEM.EPGR CDTPSRSLI~ ~~~~
Ber_2S MMRLA.ENIP SRCNL.SPMR CPMGGSIAGF ~~~~
Jre_2S MVQSA.RDLP NECGI.SSQR CEIRRSWF~~ ~~~~
Cma_2S MLQKA.RNLP SMCGI.RPQR CDF~~~~~~~ ~~~~
```

Figure 2. Aligned amino acid sequence of the putative cocoa 2S albumin with related proteins in the database. The amino acid sequence of the cocoa 2S albumin was compared with the protein sequence database of the Swiss-Prot using program FASTA (ver. 3). Pairwise comparisons were carried out by the program GAP. The multiple sequence alignments were carried out employing PILEUP. All of the software tools are available in the Wisconsin Sequence Analysis Package (ver. 8.1) from Genetics Computer Groups (GCG, Madison, WI). The residues are numbered according to the sequence of cocoa 2S albumin. The invariant residues are shown in bold face. *Ath*_2S, *Arabidopsis thaliana* 2S albumin (CAB38846); *Bra*_2S, *Brassica napus* (rape) 2S albumin (Q9S9E5); *Cam*_2S, *Capparis masaikai* 2S albumin (O04774); *Tca*_2S, *T. cacao* 2S albumin; *Ghi*_2S, *Gossypium hirsutum* (upland cotton) 2S albumin (Q39787); *Ber*_2S, *Bertholletia excelsa* (Brazil nut) 2S albumin (P04403); *Jre*_2S, *Juglans regia* (English walnut) 2S albumin (P93198); *Cma*_2S, *Cucurbita maxima* (pumpkin) 2S albumin (Q39649).

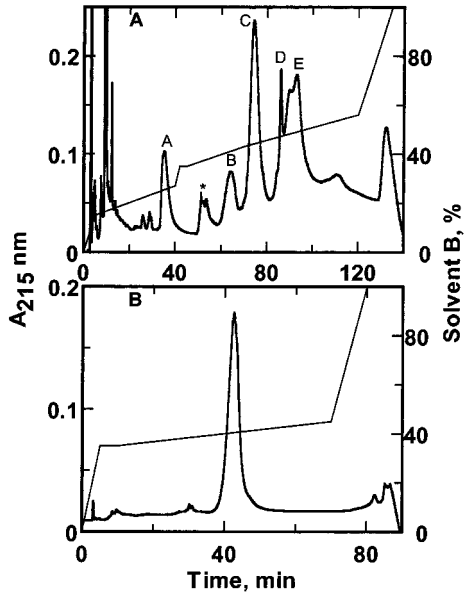

Figure 3. RP-HPLC of major cocoa polypeptides: (A) reduced and pyridinylethylated Gnd-HCl extract; (B) rechromatography of the pooled fractions of peak B from semipreparative column [column, Vydac C4, 5 μm (4.6 × 220 mm); solvent, TFA/ACN system; detection, 215 nm; flow rate, 1 mL/min]. CAP (1 g) was extracted with 10 mL of Gnd-HCl buffer (see Experimental Procedures). The extracts (1 mL) were analyzed by RP-HPLC [column, semipreparative Vydac C4, 10 μm (20 × 250 mm); solvent, TFA/ACN system; detection, 215 nm; flow rate, 4 mL/min]. Peaks: A (t_R = 37 min), vicilin fragment; B (t_R = 68 min), novel cocoa polypeptide; C (t_R = 78 min), 21 kDa storage albumin; D (t_R = 87 min), vicilin fragment; E (t_R = 91 min), 37 and 47 kDa storage vicilin. Only peaks showing polypeptide band in SDS-PAGE are denoted. The peak denoted with an asterisk is a ghost peak due to sharp changes in the gradient. For details, see Experimental Procedures.

based on the specific carbohydrate staining (kit from Bio-Rad) following SDS-PAGE analysis, is not glycosylated (data not shown). Table 2 presents the amino acid composition of the purified 9 kDa polypeptide. The cocoa polypeptide is enriched in predominantly charged residues (Glx + Asx, 30; Arg + Lys, 15).

The N-terminal sequence of the first 34 residues is RPVSKHLDSCCQQLEKLDTPCRCPGLKQAVQQQA. In addition, a second sequence of 11 amino acid residues, SKEXSCXKXI, was also detected. The initial and repetitive yields of Edman cycles were over 80 and 90%, respectively. Comparison of the N-terminal sequence against the Swiss-Prot protein sequence database shows a high degree of similarity to 2S storage proteins of cotton (50% identity; 70% similarity), Brazil nut (44% identity; 69% similarity), and sweet protein, Mabinlin (43% identity; 55% similarity), suggesting the cocoa protein to be a putative 2S albumin. The N-terminal sequence closely matches the internal sequence of the translated putative 2S albumin precursor (Figure 1), further supporting the notion that the 9 kDa cocoa polypeptide is most likely the 2S storage albumin of *Theobroma cacao*.

The primary structure of the purified polypeptide was determined by generating the tryptic peptide mass fingerprints of the reduced and pyridinylethylated polypeptide by RP-HPLC-ESI-MS. Figure 6 shows the UV elution and total ion current chromatographs of all

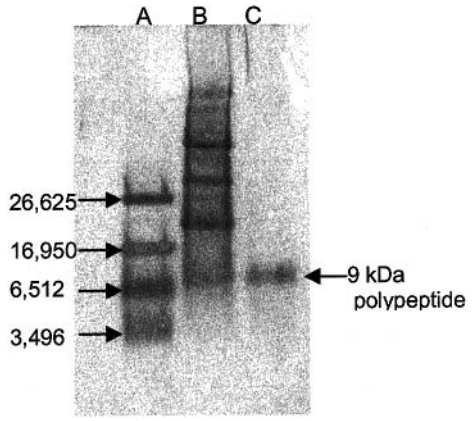

Figure 4. Tricine-SDS-PAGE analysis of the purified 9 kDa polypeptide: (lane A) molecular weight makers denoted in Da; (lane B) CAP extract (1% SDS/50 mM phosphate buffer, pH 7); (lane C) purified 9 kDa polypeptide. Analysis was carried out on 20% Tris-Tricine acrylamide gel. Ten microliter samples were diluted three times with the sample buffer in the presence of β-mecaptoethanol and heated at 100 °C for 5 min. The samples were cooled to room temperature, centrifuged, and electrophoresed at a constant 100 V until the dye reached the bottom of the gel (~2 h). The gel was stained for small peptide as directed by the supplier's instruction manual (Bio-Rad).

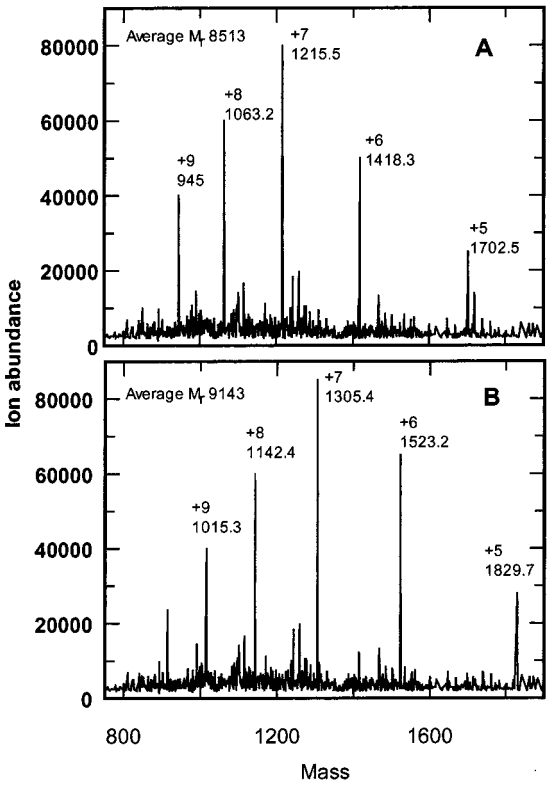

Figure 5. RP-HPLC-ESI-MS analysis of the purified 9 kDa cocoa polypeptide: (A) mass spectrum showing multiple charged ion signals (native sample; raw data from an average of 50 scans); (B) mass spectrum showing multiple charged ion signals (reduced and pyridinylethylated; raw data from an average of 50 scans). The average molecular weight was calculated following deconvolution of the charged ions by Sequest software (Finnigan-MAT).

the tryptic peptide fragments. The tryptic peptide fragments are designated "T" and are subsequently numbered on their order of elution during RP-HPLC (Figure 6). A total of 10 peptide masses are detected (Figure 6; Table 3). Most of the eluting peptides were

Table 2. Amino Acid Composition of Purified Cocoa Seed 2S Albumin

amino acid	experimentally determined	mature polypeptide[a]		deduced from cDNA	
	res/mol[b]	mol %	res/mol	mol %	res/mol
D	7	5.5	4	3.3	5
N	-	4.1	3	4	6
E	22	12.3	9	16	10.7
Q	-	11	8	8.7	13
S	2	2.7	2	9.3	14
G	4	5.5	4	6	9
H	1	1.4	1	2	3
R	5	6.8	5	5.3	8
T	5	2.7	2	4	6
A	4	2.7	2	4	6
P	4	6.8	5	4	6
C	6	8.2	6	5.3	8
Y	1	1.4	1	3.3	5
V	5	5.5	4	5.3	8
M	2	1.2	2	2.7	4
I	8	1.4	1	1.3	2
W	-	1.4	1	1.3	2
L	6	8.2	6	10	15
F	1	2.7	2	2	3
K	7	6.8	5	7.3	11
MW[c]	9200	8512		17125	

[a] Sequence position 78−150, see Figure 5. [b] Residues per mole of protein. [c] Calculated MW in Da.

detected as singly charged $[M + H]^+$ species. A comparison of the observed peptide masses against the sequence database using the software PeptideIdent shows no match to any previously identified protein. A comparison of the observed tryptic peptide masses of the mature protein against the translated amino acid sequence shows a 100% amino acid sequence match to the residues 79−147 (Figure 7). The peptide fragments containing the cysteine residues show the expected positive mass shift of 105 due to *S*-pyridinylethylation (Table 3). Every identified peptide mass was subjected to MS-MS analysis (data not shown) to determine either a complete or partial amino acid sequence to confirm its mapping to the translated amino acid sequence. The C-terminal peptide NWF was not detected. Attempts to isolate and identify peptides corresponding to the region (1−77) were not successful. Rigorous controls, for example, analysis of tryptic peptide maps generated at different substrate/enzyme ratios, RP-HPLC of the tryptic digests employing C_8 and C_{18} columns from different suppliers (Vydac, Macherey-Nagel, and Phenomenex), and extensive analysis of peptide mass fingerprint data with predictive mass peak extraction, ruled out the possibility that these peptides selectively remained undetected due to either an incomplete trypsin digestion or a chromatographic elution abnormality. The data indicate that the mature cocoa 2S albumin is post-translationally processed to yield a 73 amino acid residue polypeptide.

Maturation of the *T. cacao* 2S Albumin. The mature 2S albumins of plants such as *Brassica napus* (rapeseed) and *Cucurbita maxima* (pumpkin) are post-translationally processed to generate two subunits (*24*). In *B. napus*, this processing produces a short-chain polypeptide (4 kDa) and a long-chain polypeptide (9 kDa), which are held together by both inter- and intrachain disulfide bonds (*25*, *26*). These two peptides are produced from the 2S precursor protein by removal of the signal peptide, then removal of the 22 amino acids from the N-terminal propeptide, and removal of an internal propeptide of 28 amino acids. One or more

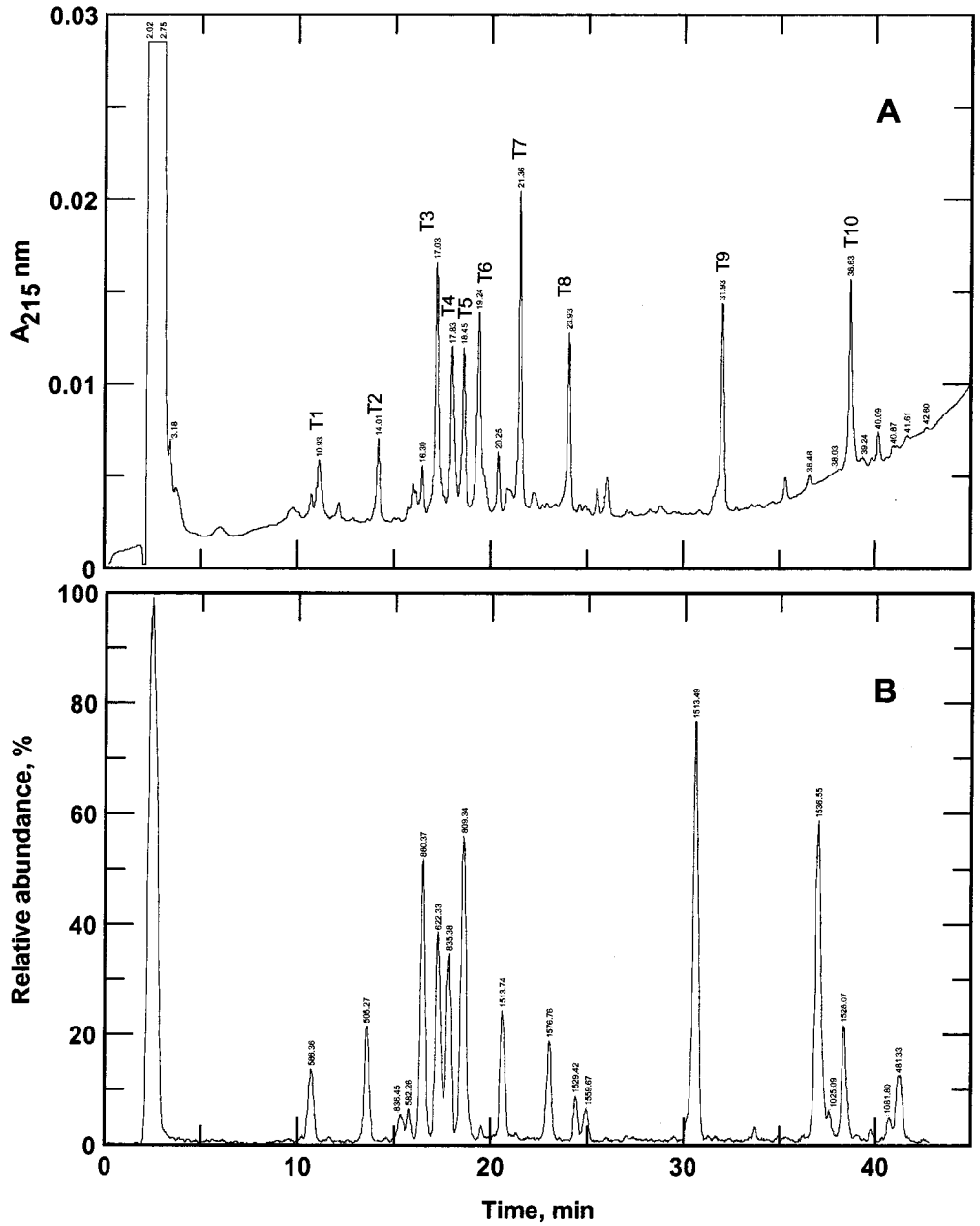

Figure 6. Tryptic peptide mass fingerprint of the purified 9 kDa cocoa polypeptide: (A) UV chromatogram with peaks denoted by retention time; (B) total ion current chromatogram with ion peaks denoted by the molecular ion mass. The RP-HPLC-ESI-MS conditions are described under Experimental Procedures.

amino acid residues are also trimmed off the 3′ end of the mature protein (*25*, *26*).

Analysis of a hydrophobicity plot (*27*) of the cocoa 9 kDa precursor protein (Figure 8) indicates that the N-terminal region of the protein encodes a distinct short hydrophobic region that is probably the signal peptide sequence. It is also interesting to note that the predicted α-helical regions (*28*) for the *T. cacao* 2S precursor show that the position of the N-terminal residue of large polypeptide fragment mapped by N-terminal sequencing (position 78, see Figure 1) has a noticeable absence of α-helical forming sequences. In fact, a similar examination of the position of the *B. napus* 2S processing sites in relation to the position of the α-helical regions indicated these sites are either on the border of predicted α-helical regions or completely outside these regions (i.e., no processing occurs in the major predicted α-helical regions of the *Brassica* 2S protein).

The details of 2S protein processing are not known for cocoa seeds. However, as stated earlier a polypeptide of ∼9 kDa has been isolated from mature cocoa seeds by RP-HPLC under denaturating conditions (6 M Gnd-HCl, reduced and pyridinylethylated), and its N-terminal sequence is identical to an internal segment of the translated sequence of the putative cocoa 2S albumin. From the position of the N-terminal amino acid of this 9 kDa peptide, it is likely that it is the cocoa equivalent of the long-chain subunit of the *B. napus* 2S protein. Further insight on the processing of the cocoa 2S albumin was obtained from the tryptic peptide mass fingerprint of the purified 9 kDa polypeptide (see Figure 6 and Table 3). Most likely this corresponds to the long-chain polypeptide (9 kDa) identified in most of the 2S seed albumins (*25*, *26*). As discussed earlier, plant 2S albumin precursors undergo processing to produce two polypeptides, a short-chain polypeptide (4 kDa) and a

Table 3. Tryptic Peptide Analysis of 9 kDa Polypeptide by LC-ESI-MS

theor av $[M + H]^+$	sequence position	tryptic peptides[a]	av obsd $[M + H]^+$
3829.042	4−39	ND[b]	
1576.735	105−118	T8	1576.7
1513.673	119−130	T7	1513.6
1439.568	55−65	ND	
1303.577	83−93	T9	1513.5[c]
836.462	76−82	ND	
775.340	135−141	T3	880.4[c]
730.366	142−147	T5	835.4[c]
724.322	66−73	ND	
704.340	94−99	T6	809.4[c]
681.287	40−45	ND	
665.362	50−54	ND	
517.280	100−104	T4	622.3[c]
505.280	131−134	T2	505.3
466.208	148−150	ND	
388.255	46−48	ND	
349.190	1−3	ND	
304.162	74−75	ND	
147.113	49−49	ND	
	NH_2 terminus[d]	T1	586.4
	unknown	T10	1536.5/1547.9[e]

[a] Figure 6. [b] Not detected. [c] Pyridinylethyl modifications at each cysteine adds a positive shift of 105 mass units. [d] Residue position 79−82. [e] $[M + H]^{2+}$ plus its sodium adduct.

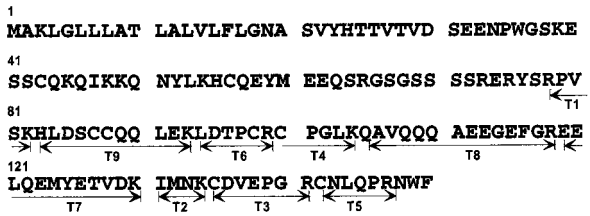

Figure 7. Mapping of the tryptic peptides of the 9 kDa cocoa polypeptide to the deduced amino acid sequence of cocoa 2S albumin precursor. The alignment of the observed peptides (see Figure 6) was based on the theoretical tryptic fragment masses from the deduced sequence generated by software PeptideMass.

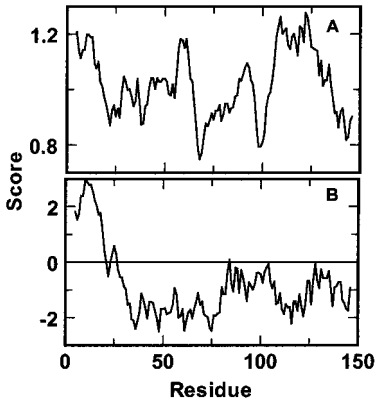

Figure 8. Predicted α-helical regions and hydrophobicity plot for the cocoa 2S albumin precursor. Secondary structure predictions (*27*, *28*) were carried out using the Wisconsin Sequence Analysis Package (ver. 8.1) from Genetics Computer Groups (GCG).

long-chain polypeptide (9 kDa) linked by both inter- and intrachain disulfide bridges, producing a heterodimeric mature 2S albumin. Attempts to isolate the 4 kDa polypeptide in cocoa seeds were not successful (data not shown). However, the second N-terminal sequence of 11 residues, despite some uncertainty, seems to correspond to residue positions 38−47 of the putative cocoa 2S protein precursor. On the basis of the known processing sites of rapeseed and pumpkin 2S proteins, one can hypothesize that cleavage at positions 30−64 in the precursor sequence should result in a polypeptide of M_r 4132. The processing at these positions will preserve all of the cysteines in this region that are presumably needed to produce a heterodimeric 2S albumin.

In conclusion, apart from the identification of certain hydrophobic amino acids, little is known about the nature and type of peptides generated from cocoa seed proteins that are important in the generation of cocoa flavor. Much attention has been focused on the determination of the degradation pattern of well-known storage proteins, with molecular weights of 47, 31, and 21 kDa. Isolation of major proteins from cocoa seed will ultimately allow us to generate specific flavor precursor pools originating from the selected proteins and enable evaluation of their impact on the overall cocoa flavor. The cocoa polypeptide described in this paper is highly expressed and shares considerable homology with the plant 2S albumins, suggesting it to be the cocoa seed storage 2S albumin. The mature 2S albumin undergoes post-translational processing to yield a polypeptide of M_r 8513 and probably a smaller polypeptide, which has yet to be identified. The mature polypeptide of M_r 8513 contains high mole percentages of Glu (23%), Cys (8%), Arg (7%), Leu (8%), Lys (7%), and Pro (7%) residues. During cocoa fermentation, the 2S albumin is extensively degraded, thus contributing to the amine group flavor precursor pool. Furthermore, recombinant expression of 2S albumin, if successful, will provide large quantities of material to study its proteolysis and resultant peptides enriching the flavor precursor pool. In a model reaction, the isolated flavor-active peptides will provide insight into the generation of peptide-based Maillard reaction products, still a poorly understood domain of organoleptically thermal reactions in foods.

ABBREVIATIONS USED

ACN, acetonitrile; CAP, cocoa acetone powder; DTT, 1,4-dithio-DL-threitol; EDTA, ethylenediaminetetraacetic acid; ESI-MS, electrospray ionization mass spectrometry; Gnd-HCl, guanidinium hydrochloride; LC-MS, liquid chromatography−mass spectrometry; PAGE, polyacrylamide gel electrophoresis; RP-HPLC, reversed-phase high-performance liquid chromatography; RT, retention time; SDS, sodium dodecyl sulfate; TIC, total ion current; TFA, trifluoroacetic acid.

ACKNOWLEDGMENT

We thank Drs. M. A. Juillerat, P. Bucheli, C. E. Hansen, N. Sprenger, and J. le Coutre for critically reading the manuscript. We give special thanks to D. Moines for excellent technical assistance.

LITERATURE CITED

(1) Mohr, W.; Roehrle, M.; Severin, Th. Uber die bildung des Kakaoaromas aus seinen Vorstifen. *Fette Seifen Anstrichsm*. **1971**, *73*, 515−521.

(2) Mohr, W.; Landschreiber, E.; Severin, Th. Zur spezifität des Kakaoaromas. *Fette Seifen Anstrichsm*. **1976**, *78*, 88−95.

(3) Voigt, J.; Biehl, B.; Kamaruddin, S.; Wazir, S. The major seed proteins of *Theobroma cacao* L. *Food Chem*. **1993**, *47*, 145−151.

(4) Voigt, J.; Biehl, B.; Heinrichs, H.; Kamaruddin, S.; Gaim-Marsonaer, G.; Hugi, A. *In-vitro* formation of cocoa-specific aroma precursors: Aroma-related peptides generated from cocoa-seed proteins by co-operation of aspartic endoprotease and a carboxypeptidase. *Food Chem*. **1994**, *49*, 173−180.
(5) Voigt, J.; Voigt, G.; Heinrichs, H.; Wrann, D.; Biehl, B. *In-vitro* studies on the proteolytic formation of the characteristic aroma precursors of fermented cocoa seeds: The significance of endoprotease specificity. *Food Chem*. **1994**, *51*, 7−14.
(6) Voigt, J.; Wrann, D.; Heinrichs, H.; Biehl, B. The proteolytic formation of essential cocoa-specific aroma precursors depends on particular chemical structure of the vicilin-class globulin of the cocoa seeds lacking in the globular storage proteins of coconuts, hazelnuts and sunflower seeds. *Food Chem*. **1994**, *51*, 197−205.
(7) Voigt, J.; Heinrichs, H.; Voigt, G.; Biehl, B. Cocoa-specific aroma precursors are generated by proteolytic digestion of the vicilin-like globulin of cocoa seeds. *Food Chem*. **1994**, *51*, 177−184.
(8) Rohn, T. A. Precursors of chocolate aroma: A comparative study of fermented and unfermented cocoa beans. *J. Food Sci*. **1964**, *29*, 456−459.
(9) Biehl, B.; Brunner, E.; Passern, D.; Quesnel, V. C.; Adomako, D. Acidification, proteolysis and flavor potential in fermenting cocoa beans. *J. Sci. Food Agric*. **1985**, *36*, 583−598.
(10) Kirchhoff, P. M.; Biehl, B.; Crone, G. Peculiarity of the accumulation of free amino acids during cocoa fermentation. *Food Chem*. **1989**, *31*, 295−311.
(11) Lerceteau, E.; Rogers, J.; Pétaird, V.; Crouzillat, D. Evolution of cocoa bean proteins during fermentation: a study by two-dimensional electrophoresis. *J. Sci. Food Agric*. **1999**, *79*, 619−625.
(12) Spencer, M. E.; Hodge, R. Cloning and sequencing of a cDNA encoding the major storage proteins of *Theobroma cacao*. *Planta* **1992**, *186*, 567−576.
(13) Spencer, M. E.; Hodge, R. Cloning and sequencing of a cDNA encoding the major albumin of *Theobroma cacao*. Identification of the protein as a member of the Kunitz protease inhibitor family. *Planta* **1991**, *183*, 528−535.
(14) Kochhar, S.; Gartenmann, K.; Juillerat, M. A. Primary structure of the abundant seed albumin of *Theobroma cacao* by mass spectrometry. *J. Agric. Food Chem*. **2000**, *48*, 5593−5599.
(15) Lundell, N.; Schreitmüller, T. Sample preparation for peptide mapping—A pharmaceutical quality-control perspective. *Anal. Biochem*. **1999**, *266*, 31−47.
(16) Lowry, O. H.; Rosebrough, N. J.; Farr, A. L.; Randall, R. J. Protein measurement with the Folin phenol reagent. *J. Biol. Chem*. **1951**, *193*, 265−275.
(17) Laemmli, U. K. Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature* **1970**, *227*, 680−685.
(18) Schägger, H.; von Jagow, G. Tricine-sodium dodecyl sulfate-polyacrylamide gel electrophoresis for the separation of proteins in the range of 1 to 100 kDa. *Anal. Biochem*. **1987**, *166*, 368−379.
(19) Heinrikson, R. I.; Meredith, S. C. Amino acid analysis by reverse-phase high-performance liquid chromatography: Precolumn derivatization with phenylisothiocyanate. *Anal. Biochem*. **1984**, *136*, 65−74.
(20) Gartenmann, K.; Kochhar, S. Short-chain peptide analysis by high-performance liquid chromatography coupled to electrospray ionization mass spectrometer after derivatization with 9-fluorenylmethyl chloroformate. *J. Agric. Food Chem*. **1999**, *47*, 5068−5071.
(21) Pearson, W. R.; Lipman, D. J. Improved tools for biological sequence analysis. *Proc. Natl. Acad. Sci. U.S.A*. **1988**, *85*, 2444−2448.
(22) Pearson, W. R. Rapid and sensitive sequence comparison with FASTP and FASTA. *Methods Enzymol*. **1990**, *183*, 63−98.
(23) Rico, M.; Bruix, M.; Gonzalez, C.; Monsalve, R. I.; Rodriguez, R. ^{1}H NMR assignment and global folding of Napin *Bnl*b, a representative 2S albumin seed protein. *Biochemistry* **1996**, *35*, 15672−15682.
(24) Hara-Nishimura, I.; Nishimura, M. Proglobulin processing enzyme in vacuoles isolated from developing pumpkin cotyledons. *Plant Physiol*. **1987**, *85*, 440−445.
(25) Ericson, M. L.; Rodin, J.; Lenman, M.; Glimelius, K.; Josefsson, L. G.; Rask, L. Structure of rapeseed 1.7S storage protein, napin, and its precursor. *J. Biol. Chem*. **1986**, *261*, 14576−14581.
(26) Bycznska, A.; Barciszewski, J. The biosynthesis, structure and properties of napin-the storage protein from rape seeds. *J. Plant. Physiol*. **1999**, *154*, 417−425.
(27) Chou, P. Y.; Fasman, G. D. Prediction of the secondary structure of proteins from their amino acid sequence. *Adv. Enzymol*. **1978**, *47*, 45−148.
(28) Kyte, J.; Doolittle, R. F. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol*. **1982**, *157*, 105−132.

Received for review April 16, 2001. Revised manuscript received July 9, 2001. Accepted July 9, 2001.

JF010497B