

Short communication

Genomic organization of peanut allergen gene, Ara h 3

Olga M. Viquez, Koffi N. Konan, Hortense W. Dodo*

*Food Biotechnology Laboratory, Department of Food and Animal Sciences, P.O. Box 1628,
Alabama A and M University, Normal AL 35762, USA*

Received 24 May 2004; accepted 9 June 2004

Available online 25 August 2004

Abstract

Type 1 hypersensitivity to peanut proteins is a well-recognized health problem. Several peanut seed storage proteins have been identified as allergens. Ara h 3, a glycinin protein, is one of the important peanut allergens. Although amino acid and cDNA sequences are available for Ara h 3, there is not information at the genomic level. The objectives of this study were to isolate, sequence, and characterize the genomic clone of peanut allergen, Ara h 3. A peanut genomic library was screened, using two [³²P] end-labeled oligonucleotide probes designed based on cDNA sequences of Ara h 3 and Ara h 4. Four positives lambda FIX II clones were obtained after four rounds of screenings. Digestion with *Sac* I resulted in two fragments of 1.5 and 10 kb hybridizing to the probes. Both fragments were subcloned into p-Bluescript vector and sequenced. The Ara h 3 gene spans 3.5 kb and consists of four exons, three introns, 5' and 3' flanking regions. The open reading frame is 2008 bp long and can encode a polypeptide of 538 amino acids residues. Sequences analogous to a TATA-box (TATAAAT), CAAT-box (AGGA), G-box (TCCTACGTGTCC) and several *cis*-elements were found in the promoter region. In the 3' downstream region, three polyadenylation signals (AATAAA) were identified.

© 2004 Elsevier Ltd. All rights reserved.

Keywords: Ara h 3; *Cis*-element; Introns; Peanut; Promoter**1. Introduction**

Food allergy is a major health problem in Western countries. Epidemiological studies have shown that the prevalence of sensitization and clinical allergy to food is increasing (Hourihane et al., 1996; Thompson and Chandra, 2002). The prevalence of peanut allergy in the US population was estimated at 0.4% for children and 0.7% for adults (Grundy et al., 2002; Sicherer, 2002). Peanut accounts for the majority of food-related allergies associated with fatal food-induced anaphylaxis (Thompson and Chandra, 2002; Ewan, 1998). Peanut allergy is increasing due to its ubiquitous use in many food products. This raises new concerns for peanut hypersensitive individuals because accidental ingestion cannot be avoided.

Peanut like many plants contains intrinsic allergens. Seven peanut allergens have been identified as seed storage proteins members of vicilin (Ara h 1), conglutin (Ara h 2, Ara h 6, Ara h 7), glycinin (Ara h 3, Ara h 4), and profilin (Ara h 5) families. Glycinins, 11S globulins, account for about 50% of the total seed globulin fraction (Beilinson et al., 2002). Ara h 3 and Ara h 4 cDNA's (GenBank accession nos. AF093541, and AF086821, respectively) have been described as two variants of the same gene. Ara h 3 is a 57-kDa protein recognized by 45% of patients with peanut hypersensitivity (Rabjohn et al., 1999). Ara h 4 is 61-kDa and recognized by more than 53% of peanut sensitive individuals (Kleber-Janke et al., 1999).

No information is available on the genomic structure and organization of peanut allergen gene, Ara h 3. We previously reported on the genomic characteristics of peanut allergen, Ara h 1 (Viquez et al., 2003) and Ara h 2 (Viquez et al., 2001). In this study, we are describing the isolation and the structural characterization of the genomic DNA of peanut allergen gene Ara h 3.

* Corresponding author. Tel.: +1 256 372 4171; fax: +1 256 372 5432.
E-mail address: aamhwd01@aamu.edu (H.W. Dodo).

2. Materials and methods

2.1. Isolation of *Ara h 3* genomic clones

The genomic library used in this research has been previously described (Viquez et al., 2001). For library screening, two specific 100 bp probes were designed based on conserved regions of published cDNA sequences of peanut allergens, *Ara h 3* and *Ara h 4*. The 5' flanking oligonucleotide used as probe (5'-gggcggttacattgagacttggaaacccaacaaccaggaggttcgaatgcgccggcgctcgccctctctcgcttagtctccgcaacgccttcgtagg-3') corresponds to nucleotides 78–177 of the *Ara h 3* cDNA and 147–246 of the *Ara h 4* cDNA. The 3' flanking oligonucleotide used as probe (5'-gaggggacgggctcagctgcaagtcgtggacagcaacggcaacagagtgtacgacgaggagctcaagaggggtcacgtgctgtggtgccacagaacttc-3') corresponds to nucleotides 1191–1290 of the *Ara h 3* cDNA and 1260–1359 of the *Ara h 4* cDNA. The probes were [³²P] end-labeled as described by Fermentas Inc. (Hanover, MD). About 300,000 plaques of the recombinant lambda FIX II peanut genomic library were screened, using plaque hybridization protocol as previously described by Viquez et al. (2003). Four positive-hybridizing plaques were isolated after four rounds of screening. Lambda DNA from each plaque was extracted, using a lambda kit from Qiagen Inc. (Valencia, CA).

2.2. Subcloning and sequencing

Lambda DNA from putative positive clones was digested with *Sac* I and *Xba* I. These enzymes were selected because of the presence of their restriction sites within the multiple cloning region of the Lambda FIX II vector. Digested DNA was electrophoresed on a 0.65% agarose gel and transferred by capillary action to Hybond N + nylon membranes according to the manufacturer's instructions (Amersham Pharmacia). Prehybridization (30 min) and hybridization (2 h) were performed in HyperHyb buffer (ResGen Inc., Huntsville, AL) at high-stringency conditions (65 °C). The membranes were washed 3 times for 15 min in 0.1 × SSC + 0.1% SDS at 65 °C. The final wash was in 1 × SSC at room temperature, and membranes were exposed to X-ray film (Kodak, Biomax™ MS) at room temperature overnight. Analysis of the X-ray film revealed that two *Sac* I and one *Xba* I-fragments hybridized to both probes. The two *Sac* I-fragments of approximately 1.5 and 10 kb in size were isolated and ligated separately into the *Sac* I site of a p-Bluescript vector (Stratagene, La Jolla, CA) for DNA sequencing.

The nucleotide sequences were determined by the ABI PRISM (Dye Terminator Cycle Sequencing Ready Reaction kit, using AmpliTaq DNA Polymerase, FS) at ResGen Inc. (Huntsville, AL).

2.3. Sequence analysis

Sequence analysis was performed, using the BLAST and BLAST 2 sequences tools (Altschul et al., 1997; Tatusova and

Madden, 1999a,b). Amino acid sequence alignments were undertaken, using the ClustalW program. Signal peptide analysis was performed as described by Nielsen et al. (1997). Cis-elements were identified, using MatInspector (Quandt et al., 1995).

3. Results and discussion

3.1. Nucleotide sequence analysis of the open reading frame (ORF) of the *Ara h 3* gene

Four positive clones were isolated from the peanut genomic library after screening with two [³²P] end-labeled synthetic probes, and sequenced. The sequences were identical and revealed a full-length genomic clone (GenBank accession no. AF10854) that spans 3828 bp in size with an ORF of 2008 bp that comprises 1614 bp of coding regions and three introns of 594 bp total.

The ORF starts with an ATG codon (position +1) located within the first exon and ends with a TAA codon (position +2209) within exon 4 (Fig. 1). The nucleotide sequences surrounding the start codon match the plant consensus for initiation of translation for plant genes (ANNATGGC) (Grierson and Covey, 1988; Joshi et al., 1997; Joshi et al., 1997; Lukaszewicz et al., 2000).

Alignment with the published cDNA sequences (Kleber-Janke et al., 1999; Rabjohn et al., 1999) showed that the *Ara h 3* gene has four exons (289, 293, 645, and 387 bp) and three introns (339, 87, and 168 bp) (Fig. 1). Similar structures have been identified in other glycinin genes in soybean, such as Gy1, Gy2, Gy3, Gy4, and Gy5 (Nielsen et al., 1989).

Plant introns are known to be non-coding regions rich in adenosine (A), and thymidine (T) (Csank et al., 1990; Goodall and Filiowicz, 1989); with splice junctions at the left or 5' border (5'-GT-3') and right or 3' border (5'-AG-3') (McCullough and Schuler, 1993). All exon–intron junction sites present in the *Ara h 3* gene conformed to the GT/AG rule (Table 1) as shown in other plant genes including the major peanut allergen gene, *Ara h 1* (Viquez et al., 2003). *Ara h 3* introns are rich in A + T content (Table 1). The exon/intron boundaries of the identified gene are more highly conserved at the 5' border (CAAGGT).

The presence of functional introns, exonic-border sequences and the strength of the promoter have been associated with higher expression of mRNA (Lorković et al., 2000). Whether the presence of introns enhances the production of the *Ara h 3* protein is currently unknown. However, the fact that 3 introns were found in the genomic sequence of another major peanut allergen gene, *Ara h 1* (Viquez et al., 2003), could suggest that they might play a critical role in the expression of these two major peanut allergens.

The fourth and last exon contains the termination codon (TAA), which is conserved between the *Ara h 3* and *Ara h 4* cDNA sequences.

Fig. 1. Nucleotide and deduced amino acid sequences of the genomic clone of the gene encoding peanut allergen Ara h 3 (GenBank accession no. AF10854). Uppercase letters indicate exon sequences and lowercase letters represent introns. The putative CAAT box, TATA box, start codon (ATG), stop codon (TAA) and polyadenylation signals (AATAAAA) are underlined. The putative signal peptide is italicized and underlined. The deduced amino acid sequence is below the corresponding DNA sequence.

Table 1
Exon–intron splice junctions of the Ara h 3 gene

Exon no.	Exon size (bp)	Sequence at exon–intron junction									Intron no.	Intron size (bp)
		5'splice donor					3'splice acceptor					
1	289	ATC ILe	CAG Gln	CAA Gln	G	gttattactactttt. . .tatgatgacatacag	GA Gly	AGG Arg	GGA Gly	I	339	
2	293	TTC Phe	CCC Pro	AGG Arg		gtatatatatgtttc. . .aataatacataacag	AGA Arg	TTC Phe	AAT Asn	II	87	
3	645	CTC Leu	TAC Tyr	AGG Arg		gtttgtactttctct. . .gaaatgtgattgcag	AAT Asn	TTC Ala	AAT Leu	III	168	
4	37											

3.2. Nucleotide sequences analysis of the untranslated regions (UTR) of the Ara h 3 gene

3.2.1. 5' Upstream promoter region

About 660 bp of the promoter region were sequenced. A putative TATA box was found at position –64. The TATA box (TATAAAT) matches the established consensus sequence for plant genes (TATAT/AA_{1–3}C/T) (Grierson and Covey, 1988). A CAAT box (AGGA) is located at position –146 bp from the start codon or –78 bp from the TATA box (Fig. 2). This box is 100% identical to the consensus sequence for plant genes (AGGA). Previous studies have shown that these two elements are important for RNA polymerase II binding and for the initiation of transcription (McKnight and Kingsbury, 1982). Another regulatory sequence was a G-box (TCCTACGTGTCC) at position –88. This box is important for promoter activity and has been associated as a *cis*-acting element playing a role in gene's transcription (De Vetten and Ferl, 1994). Eleven putative *cis*-elements were located within the Ara h 3 promoter (Fig. 2).

3.2.2. 3' Downstream region

About 954 bp of the 3'UTR was sequenced. Three polyadenylation signals (AATAAA) were identified at positions 2434, 2768, and 2899. The first polyadenylation signal is located 222 bp from the stop codon. This position is similar to the polyadenylation signal found in Ara h 4 cDNA (Kleber-Janke et al., 1999). In addition, the 3' flanking region has an 85% nucleotide identity to the 3'UTR of the published Ara h 4 cDNA sequence.

3.3. Analysis of the deduced amino acids

The deduced protein of 538 amino acids residues has a calculated molecular mass of 61.7 kDa, which is comparable to the molecular mass 57 and 61 kDa of Ara h 3 and Ara h 4 proteins, respectively (Kleber-Janke et al., 1999; Rabjohn et al., 1999).

A computer analysis of the coding region reveals the presence of an N-terminal transit peptide of 20 amino acids (MGKLLALSVCFCFLVLGASS) enriched with hydrophobic amino acids that make up 55% of all residues. Leucine is the most common amino acid (25%). The leader sequence is

important for protein translocation to the vacuole of the cell (Martoglio and Dobberstein, 1998). This amino acid signal peptide is present in the Ara h 4 cDNA sequence; therefore, one can postulate that the transit peptide was not removed after post-transcription for the Ara h 4 protein, and this accounts for apparent differences between Ara h 3 and Ara h 4 which are in fact variant of the same gene.

Analysis of the deduced amino acid sequence showed that the encoded protein has 93 and 91% amino acid identity with peanut allergens Ara h 3 and Ara h 4 cDNA's, respectively. There is 96 and 95% homology at the nucleotide level with Ara h 3 and Ara h 4 cDNAs, respectively. This high homology suggests that these proteins are potentially the same and that apparent sequence differences are the result of partial proteins, sequences misreading and/or sequencing errors.

Comparison with other seed storage proteins showed that the deduced amino acid sequence of Ara h 3 has conserved regions with glycinins in the following species: *Arachis hypogaea* L. (94%) and *Glycine Max* (60%); and legumins from the *Vicia* (50%). These similarities highlight the potential for cross-reactions between these proteins and peanut allergens which can be clinically significant.

Analysis of the primary amino acid sequence of Ara h 3 also provides information on the presence of two main subunits of the cupin superfamily. The cupin superfamily comprises a family of proteins with or without enzymatic activities possessing a common beta-barrel structure that can be divided into two groups based on their ancient domain content, the simple structure with a single domain and the complex structure with double domains (Dunwell et al., 2000). The presence of the two cupin domains and their respective motifs confirms that the protein encoded by the Ara h 3 genomic clone is a member of the cupin superfamily. Several allergenic seed storage proteins have been identified as members of this family including 7S globulins of soybean (beta-conglycinin), peanut (conarachin; Ara h 1), walnut (Jug r 2) and lentil, and the 11S globulins of peanut (arachin; Ara h 3), and soybean (glycinin) (Mills et al., 2001). Conserved regions with glycinins and legumins include amino acids 37–204 and 369–515 of the Ara h 3 deduced amino acid sequence. These conserved regions correspond to the two cupin domains previously identified for the Ara h 3 protein (Mills et al., 2001), and suggests an evolution from a common ancestor.

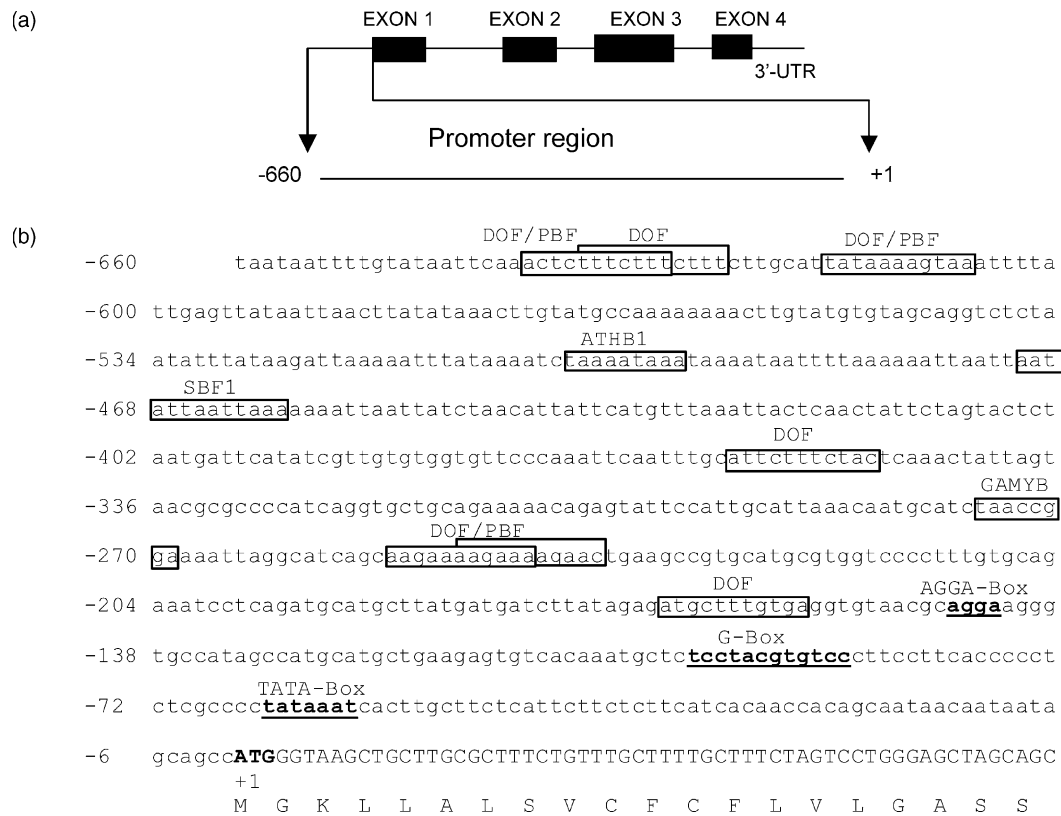


Fig. 2. (a) Structure of the peanut Ara h 3 gene. Thick boxes indicate exons and a line between them mark introns. (b) Enlargement of the promoter region showing several putative regulatory elements. The initiation start site (ATG) for the Ara h 3 is bolded and shown as +1. Partial nucleotide of the first exon with its corresponding amino acid is presented. The putative TATA-box, AGGA-box and G-box are bolded and underlined. Putative binding sites for transcription factors are boxed according to their position.

Acknowledgements

We gratefully acknowledge the assistance of ResGen (Huntsville, AL, USA). We thank Dr. Albert Abbott (Department of Biological Sciences, Clemson University, Clemson, SC, USA), for providing the peanut Genomic library. This work was funded by a Grant from USDA.

References

- Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* 25, 3389–3402.
- Beilinson, V., Chen, Z., Shoemaker, R.C., Fischer, L., Goldberg, B., Nielsen, C., 2002. Genomic organization of glycinin genes in soybean. *Theor. Appl. Genet.* 104, 1132–1140.
- Csank, C., Taylor, F.M., Martindale, D.W., 1990. Nuclear pre-mRNA introns: analysis and comparison of sequences from *Tetrahymena thermophila* and other eukaryotes. *Nucleic Acids Res.* 18, 5133–5141.
- De Vetten, N.C., Ferl, R., 1994. Transcriptional regulation of environmentally inducible genes in plants by an evolutionary conserved family of G-box binding factors. *Int. J. Biochem.* 26 (9), 1055–1068.
- Dunwell, J.M., Khuri, S., Gane, P.J., 2000. Microbial relatives of the seed storage proteins of higher plants: conservation of structure and diversification of function during evolution of the cupin superfamily. *Microbiol. Mol. Biol. Rev.* 64, 153–179.
- Ewan, P.W., 1998. Anaphylaxis. *Br. Med. J.* 316, 1442–1445.
- Goodall, G.J., Filiowicz, W., 1989. The AU-rich sequences present in the introns of plant nuclear pre-mRNAs are required for splicing. *Cell* 58, 473–483.
- Grierson, D., Covey, S.N., 1988. Structure and expression of nuclear genes. In: *Plant Molecular Biology*. A Blackie and Son Ltd., Chapman and Hall Publishers, New York, pp. 23–46.
- Grundy, J., Matthews, S., Bateman, B., Dean, T., Arshad, S., 2002. Rising prevalence of allergy to peanut in children: Data from 2 sequential cohorts. *J. Allergy Clin. Imm.* 110, 784–789.
- Houihane, J.O., Dean, T.P., Warner, H.O., 1996. Peanut allergy in relation to heredity, maternal diet, and other atopic diseases: results of a questionnaire survey, skin prick testing, and food challenges. *Br. Med. J.* 313, 518–521.
- Joshi, C.P., Zhou, H., Huang, X., Chiang, V.L., 1997. Context sequences of translation initiation codon in plants. *Plant Mol. Biol.* 35, 993–1001.
- Kleber-Janke, T., Cramer, R., Appenzeller, U., Schlaak, M., Becker, W.M., 1999. Selective cloning of peanut allergens, including profilin and 2S albumins, by phage display technology. *Int. Arch. Allergy Immunol.* 119, 265–274.
- Lorković, Z.L., Wiczorek, D.A., Lambermon, M.H.L., Filipowicz, W., 2000. Pre-mRNA splicing in higher plants. *Trends Plant Sci.* 5, 160–167.
- Lukaszewicz, M., Feuermann, M., Jérôme, B., Stas, A., Boutry, M., 2000. In vivo evaluation of the context sequence of the translation initiation codon in plants. *Plant Sci.* 154, 89–98.

- Martoglio, B., Dobberstein, B., 1998. Signal sequences: more than just greasy peptides. *Trends Cell Biol.* 8, 410–415.
- McCullough, A.J., Schuler, M.A., 1993. AU-rich elements affect pre-mRNA 5' splice site selection in *Drosophila melanogaster*. *Mol. Cell Biol.* 13, 7689–7697.
- McKnight, S.L., Kingsbury, R., 1982. Transcriptional control signals of a eukaryotic protein-coding gene. *Science* 217, 316–324.
- Mills, E.N., Jenkins, J., Marigheto, N., Belton, P.S., Gunning, A.P., Morris, V.J., 2001. Allergens of the cupin superfamily. *Biochem. Soc. Trans.* 30, 925–929.
- Nielsen, C.H., Dickinson, C.D., Cho, T.J., Thank, V.H., Scallan, B.J., Fisher, R.L., Thomas, L.S., Drews, G.N., Golberg, R.B., 1989. Characterization of the glycinin gene family in soybean. *The Plant Cell* 1, 313–328.
- Nielsen, H., Engelbrecht, J., Brunak, S., von Heijne, G., 1997. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng.* 10, 1–6.
- Quandt, K., Frech, K., Karas, H., Wingender, E., Weerner, T., 1995. MatInd and MatInspector-New fast and versatile tools for detection of consensus matches in nucleotides sequence data. *Nuc. Acids Res.* 23, 4878–4884.
- Rabjohn, P., Helm, E.M., Stanley, J.S., West, C.M., Sampson, H.A., Burks, A.W., Bannon, G.A., 1999. Molecular cloning and Epitope analysis of the peanut allergen Ara h 3. *J. Allergy Clin. Imm.* 103, 535–542.
- Sicherer, S., 2002. Clinical update on peanut allergy. *Ann. Allergy Asthma Immunol.* 88, 350–361.
- Tatusova, T.A., Madden, T.L., 1999a. *FEMS Microbiol. Lett.* 174 (2), 247–250.
- Tatusova, T.A., Madden, T.L., 1999b. *FEMS Microbiol. Lett.* 177 (1), 187–188.
- Thompson, K., Chandra, R.K., 2002. The management and prevention of food anaphylaxis. *Nutr. Res.* 22, 89–110.
- Viquez, O.M., Koffi, K.N., Dodo, H.W., 2003. Structure and organization of the genomic clone of a major peanut allergen gene, Ara h 1. *Mol. Immunol.* 40, 565–571.
- Viquez, O.M., Summer, C.G., Dodo, H.W., 2001. Isolation and molecular characterization of the first genomic clone of a major peanut allergen, Ara h 2. *J. Allergy Clin. Immunol.* 107, 713–717.