



PERGAMON

Available online at www.sciencedirect.com

SCIENCE @ DIRECT®

Molecular Immunology 40 (2003) 565–571

**Molecular
Immunology**www.elsevier.com/locate/molimm

Structure and organization of the genomic clone of a major peanut allergen gene, *Ara h 1*

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

Food Biotechnology Laboratory, Department of Food and Animal Sciences, Alabama A&M University, Normal, AL 35762, USA

Received 11 July 2003; received in revised form 11 July 2003; accepted 2 September 2003

Abstract

Peanut is one of the most allergenic foods. It contains multiple seed storage proteins identified as allergens, which are responsible for triggering IgE-mediated allergic reactions. *Ara h 1* is a major peanut allergen recognized by over 90% of peanut sensitive population. The objectives of this study were to isolate, sequence, and determine the structure and organization of at least one genomic clone encoding *Ara h 1*. Two 100 bp oligonucleotides were synthesized and used as probes to screen a peanut genomic library constructed in a Lambda FIX II vector. After three rounds of screening, four putative positive clones were selected and their DNA digested with *SacI*. A unique 12–13 kb insert fragment was released, confirmed positive by Southern hybridization, subcloned into a pBluescript vector, and sequenced. Sequence analysis revealed a full-length *Ara h 1* gene of 4447 bp with four exons of 721, 176, 81 and 903 bp and three introns of 71, 249 and 74 bp. The deduced amino acid encodes a protein of 626 residues that is identical to the *Ara h 1* cDNA clone P41b. Several well characterized elements for promoter strength were found in the promoter region of *Ara h 1* and include two TATA-boxes (TATATAAATA and TTATATATAT) at positions –89 and –348, respectively; a CAAT-box (CAAT) at position –133, a GC-box (CGGGACCGGGCCGG GCCTTCGGGCGGGCCGGGT) at position –475, two G-boxes (TAACACGTACAC and ATG-GACGTGAAA) at positions –264 and –1808, respectively; two RY elements (CATGCAC and CATGCAT) at positions –235 and –278, respectively; and other *cis*-element sequences. In the 3' UTR, a poly-A signal (AATAAA) was found at +2350, two additional stop codons (TAA) at +2303 and +2306, and TTTG/CTA/G motifs. Three introns and a potentially strong promoter could explain the high expression of the *Ara h 1* gene. Amino acid sequence comparisons revealed high sequence similarity with other plant vicilins, member of the cupin superfamily.

© 2003 Elsevier Ltd. All rights reserved.

Keywords: A-box; CAAT-box; Cupin; Exons; Genomic library; G-box; GC-box; *cis*-Elements; Introns; Peanut allergy; Promoter; RY-element; Seed storage protein; TATA-box; Vicilin

1. Introduction

Peanut (*Arachis hypogaea* L.) is a source of some of the most severe food allergens. It is estimated that three million Americans (Sicherer et al., 1999) and 1 in 200 of the UK population (Emmett et al., 1999) are allergic to peanut and tree nuts, and approximately 125 of them die each year in the USA from accidental ingestion (Burks et al., 1999). However, the only therapeutic option available is the avoidance of peanut.

Peanut allergens are seed storage proteins belonging to two major globulin families: arachin (legumin) and conarachin (vicilin). They accumulate to high levels in the developing seeds (Kang and Gallo-Meagher, 2002) and might serve as source of nitrogen and amino acid during

seed germination (Higgins, 1984). Multiple peanut allergens have been identified at the amino acid and cDNA levels. They are: *Ara h 1* (vicilin), *Ara h 2* (conglutin), *Ara h 3* (glycinin), *Ara h 4* (glycinin), *Ara h 5* (profilin), *Ara h 6* (conglutin), and *Ara h 7* (conglutin) (Burks et al., 1991, 1995; De Jong et al., 1998; Kleber-Janke et al., 1999; Rabjohn et al., 1999; Stanley et al., 1997). *Ara h 1* and *Ara h 2* are classified as major allergens because they are recognized by serum IgE in over 90% of peanut-hypersensitive patients. *Ara h 1* is a member of the vicilin family with a molecular weight of 63.5 kD (Burks et al., 1991). It is resistant to heat (Maleki et al., 2000), stable to digestion, and highly structured at the secondary level (Koppelman et al., 1999). Two cDNA clones have been identified as isoforms of *Ara h 1* (GenBank accession nos. L34402, L38853).

To date, only peanut allergen *Ara h 2* has been sequenced and characterized at the genomic level (Viquez et al., 2001).

* Corresponding author. Tel.: +1-256-372-4173; fax: +1-256-372-5432.
E-mail address: oviquez@aamu.edu (O.M. Viquez).

The understanding of the structure and organization of this and other allergen genes will help in optimizing our overall goal of silencing allergen genes for the development of an hypoallergenic transgenic peanut.

In our current study, we describe the isolation and cloning of a gene encoding the major peanut allergen *Ara h 1*. We have determined the structure of the gene and identified several transcription factor motifs which may play a role in regulating the expression of the *Ara h 1* gene.

2. Materials and methods

2.1. Construction of synthetic oligonucleotides

Two synthetic oligonucleotides were constructed based on the published cDNA sequences of peanut allergen *Ara h 1* clones P41b and P17 (GenBank accession nos. L34402, L38853). The first oligo, *Arah1*(1): (5'-CCGCCAAC-CCGGAGACTACGATGATGACCGCCGTCACCCCGA-AGAGAGGAAGGAGGCCGATGGGGACCAGCTGGAC-CGAGGGAGCGTGA-3') was deduced from the 5' end and corresponds to nucleotides 331–420 of *Ara h 1* cDNA clone P41b and 266–356 of cDNA clone P17. The second oligonucleotide, *Arah1*(2): (5'-TTCAACTCAAAGGCCATGGTTA-TCGTCGTCGTCACAAAGGAAGTGGAACTTGAAC-TCGTGGCTGTAAAGAAAAGAGCAACAACAGAGGGG-ACGGCGGG-3') was deduced from the 3' end and corresponds to nucleotides 1391–1490 of *Ara h 1* cDNA clone P41b and 1320–1419 of cDNA clone P17. About 10 pmol of each purified oligonucleotide were [³²P] end-labeled as described by Fermentas Inc. (Hanover, MD, USA) and used as probes to screen a peanut genomic library.

2.2. Isolation of peanut *Ara h 1* genomic clones

A peanut genomic library constructed in a Lambda FIX II vector was screened with both [³²P]-labeled probes using plaque hybridization protocol (Sambrook et al., 1989). In brief, DNA from phage plaques were transferred to duplicate sets of nylon membranes (Perkin-Elmer Life and Analytical Sciences Inc., Albany Street Boston, MA, USA) by a plaque lifting method. The phage DNAs on the membrane were denatured, neutralized, and cross-linked. One set of membranes was used for each probe. Low-stringency prehybridization (42 °C for 3 h) was conducted in a solution of 20× SSPE (30%) + 10% SDS (5%) + 100× Denhardt's solution (5%) + 5% sodium pyrophosphate (Na₄P₂O₄·10H₂O) (1%) + 10 mg/ml salmon sperm DNA (1%). During hybridization, individual labeled probes were added to a solution of 20× SSPE (30%) + 10% SDS (5%) + 100× Denhardt's solution (1%) + 5% sodium pyrophosphate (Na₄P₂O₄·10H₂O) (1%). After overnight incubation, the membranes were washed at 42 °C for 15 min in 6× SSPE + 0.1% SDS, followed by 15 min in 2× SSPE + 0.1% SDS and 15 min in 0.2×

SSPE + 0.1% SDS, and exposed to X-ray film (Kodak, Biomax TM, MS, USA). Two additional screenings were performed as described above except that the temperature was raised to 50 °C. At the end of the first screening, 24 putative positive clones hybridized to probe *Arah1*(1) and 16 with probe *Arah1*(2). However, only six clones were simultaneously positive for both probes and only those were selected for a second screening. At the end of the second screening, 16 clones hybridized to both probes and 8 of those were confirmed positives by a third screening.

DNA was extracted using the Lambda Kit protocol (Qia-gen Inc., Valencia, CA, USA) according to manufacturer's instructions. Purified DNA was digested with *SacI* at 37 °C overnight (Sambrook et al., 1989) and electrophoresed on a 0.65% agarose gel. One insert fragment of 12–13 kb was released, and a Southern hybridization was performed with each of the [³²P]-labeled probes using high-stringency conditions at 65 °C in HyperHyb buffer (ResGen Inc., Huntsville, AL, USA) as previously described (Viquez et al., 2001).

The positive fragment was subcloned into pBluescript II SK(+/-) vector (Stratagene, La Jolla, CA, USA). The ligation reaction was performed at 4 °C for three days and transformed into one shot TOP10 cells (Invitrogen Co., Carlsbad, CA, USA). Positive colonies were selected by blue-white color selection and the presence of the 12–13 kb *Ara h 1* insert fragment in the plasmid was confirmed after digestion with *SacI* and Southern hybridization was performed using both probes as described above.

2.3. Sequencing and sequence analysis

Positive pBluescript clones were purified as previously described (Viquez et al., 2001). DNA were sequenced in both sense and antisense orientations using standard sequencing primers, T3 and T7, as well as primers specific to the *Ara h 1* sequence. The sequencing was performed with the ABI PRISM (Dye Terminator Cycle Sequencing Ready Reaction kit using AmpliTaq DNA Polymerase, FS) at ResGen Inc. (Huntsville, AL, USA).

Nucleotide sequence analysis and homology searches were performed using the BLAST and BLAST 2 sequence tools (Atschul et al., 1997; Tatusova and Madden, 1999). Determination of the protein leader sequence was performed as previously described by Grierson and Covey (1988) and Nielsen et al. (1997). Amino acid sequences were performed using DNASTAR and MEGA version 2.1 programs (DNASTAR, Madison, WI, USA). The Reverse Position Specific Blast (RPS-BLAST) program (Atschul et al., 1997) was used to determine the presence of the conserved cupin domain within *Ara h 1*. The *Ara h 1* promoter region was analyzed for putative *cis*-elements by matrice analysis using the Transfac database and the Matinspector Program (Biological Database, Braunschweig, Germany) (Quandt et al., 1995).

3. Results and discussion

3.1. Structure of the peanut allergen *Ara h 1* gene

We report here, for the first time, the isolation, sequencing, and structural organization of a genomic clone encoding the major peanut allergen, *Ara h 1*. Eight putative genomic clones were isolated, one was sequenced and revealed a full-length genomic clone (GenBank accession no. AF432231) of 4477 bp long.

Based on the DNA sequence information, the *Ara h 1* has an open reading frame of four exons of 721, 176, 81 and 903 bp, and three introns of 71, 249 and 74 bp. Exon 1 contains the coding information for the first 240 amino acids of the protein, including the putative signal peptide of 27 amino acid residues (Fig. 1). Exon 2 codes for 59, exon 3 for 27 and exon 4 for 300 amino acid residues. When the exon sequences are read continuously, they code for a polypeptide of 626 amino acid residues.

The coding region of *Ara h 1* is comprised of 1878 bp and revealed 100% and 97.8% nucleotide sequence homologies with *Ara h 1* cDNA clones P41b and P17, respectively. Therefore, it is speculated that the genomic *Ara h 1* clone isolated in this study is the one transcribed to produce the mRNA for cDNA clone P41b.

The three introns identified in the *Ara h 1* gene are of the class III like most plant introns and are rich in AT-rich regions (Grierson and Covey, 1988). They are 73.2, 73.5 and 79.7% A + T rich for the first, second and third intron, respectively (Table 1). All the splice junctions between exons and introns of the *Ara h 1* respect the AG-GT consensus sequence rule (Lorković et al., 2000; Grierson and Covey, 1988). The consensus sequences at the 5' and 3' intron border of legume genes are **GT**(AG)AG(CT) and T(TCG)**CAG**, respectively (Häger et al., 1996). Sequences at the 5' border of the three *Ara h 1* introns match the consensus sequence, while only intron 1 and 2 are conserved at the 3' border and match the consensus sequence (Table 1).

3.2. Promoter region of the *Ara h 1* gene

The 5'-flanking promoter region of the *Ara h 1* gene is 1926 bp long and contains several conserved sequences motifs important for gene expression. Two putative TATA boxes (TATATAAATA and TATATATATAT) were identified at positions –98 and –357, respectively. The first box is 100% identical to the consensus sequence for plant genes

(TATAT/AA_{1–3}C/T) (Grierson and Covey, 1988) (Fig. 2). Most TATA boxes in plants are generally positioned around 16–54 bp upstream of the 'cap site'. However, it is not unusual to observe plant genes such as the peanut lectin gene (Law, 2000) with a TATA-box at position –95, and the soybean actin gene with a TATA box at –99 bp (Grierson and Covey, 1988). TATA-boxes are important for the positioning of RNA polymerase II and to initiate transcription of protein producing genes.

A putative CAAT-box was located at position –137 from the start codon or 34 bp from the first TATA-box (Fig. 2). Its sequence is in agreement with plant consensus (CCAAT/AGGA) (Grierson and Covey, 1988). A GC-box was identified at position –475 and characterized by a G + C-rich (88%) region of 34 nucleotides. This is of particular interest because of the similarity between this region with the monocot-specific promoter (MSP) element. Thus, the MSP motif is not a specificity of monocots, but it is also found in dicot genes. Several copies of this motif, consensus 5'-RGGGCR-3', have been associated with the transcription of the rice *U3* gene (Qu et al., 1996) and small nuclear RNA genes (Connelly et al., 1994). Inspection of this motif (5'-CGGGACCGGGCCGGGCCTTCGGGCCGGGCCGGGT-3') reveals six repeated sequences of the MSP consensus sequence. We believe that the GC-box found in the *Ara h 1* promoter is likely to be a binding site of a regulatory element and with similar function to those found in monocot plant genes. Two G-boxes (TAACACGTACAC and ATG-GACGTGAAA) are located at positions –264 and –1819, respectively. The G- and GC-boxes are the most widely *cis*-acting elements in eukaryotic genes used for transcription (De Vetten and Ferl, 1994). Two GATA binding motifs, both with consensus GATGATAT were identified at positions –439 and –1329, respectively. GATA binding motifs, consensus T/AGATAA/G, have been associated with the tissue-specific expression of several genes (Jepeal et al., 2003). An A-box-like sequence, consensus TACGCTT, was identified at position –1013. DNA sequence motifs containing an ACGT core specially in A-, C- and G-boxes have been identified as binding sites for plant basic leucine zipper (bZIP) proteins, which are known to play an important role in the efficiency of gene transcription (Izawa et al., 1993; Foster et al., 1994). Moreover, AT-rich sequences were identified within the promoter region at positions (–758 to –747; –824 to –813 and –1122 to –1113). AT-rich regions have been reported to enhance gene expression (Strömvik et al., 1999), therefore these regions

Table 1

Size, percent A + T and sequence characteristics of the three introns of the peanut *Ara h 1* genomic clone

Intron	Intron/exon boundaries	Size (bp)	A + T (%)	Nucleotide position
1	CA AG /GTATCAAAT...TTGGATTGC AG /G	71	73.2	812–882
2	T AG /GTACCTCTT...TGCTTC AG /G	249	73.5	883–1058
3	CA AT /GTAAGCAAA...TATATACAT AG /G	74	79.7	1308–1387

At the intron/exon boundaries the splice-sites are in bold and the conserved sequences are highlighted in gray.

```

-81 cctccctccatcatttcaccatccacactcataataatcatatatattcatcaatcatctatataagtagtagcaggagca
1  ATGAGAGGGGAGGGTTTCTCCACTGATGCTGTTGCTAGGGATCCTTGTCTGGCTTCAGTTTCTGCAACGCATGCCAAGTCA
   M R G R V S P L M L L L G I L V L A S V S A T H A K S
81  TCACCTTACCAGAAAGAAACAGAGAACCCTGCGCCAGAGGTGCCTCCAGAGTTGTCAACAGGAACCGGATGACTTGAA
   S P Y Q K K T E N P C A Q R C L Q S C Q Q E P D D L K
163  CAAAAGGCATGCGAGTCTCGCTGCACCAAGCTCGAGTATGATCCTCGTTGTGTCTATGATCCTCGAGGACACACTGGCACC
   Q K A C E S R C T K L E Y D P R C V Y D P R G H T G T
244  ACCAACCAACGTTCCCTCCAGGGGAGCGGACAGTGGCCGCCAACCCGGAGACTACGATGATGACCGCGTCAACCCCGA
   T N Q R S P P G E R T R G R Q P G D Y D D D R R Q P R
325  AGAGAGGAAGGAGGCGGATGGGGACAGCTGGACCGGAGCGTGAAAGAGAAGAAGACTGGAGACAACCAAGAGAAGAT
   R E E G G R W G P A G P R E R E E D W R Q P R E D
406  TGGAGGCGACCAAGTCATCAGCAGCCACGGAATAAGGCCGAAGGAAGAGAAGGAGAACAAGAGTGGGGAACACCAGGT
   W R R P S H Q Q P R K I R P E G R E G E Q E W G T P G
487  AGCCATGTGAGGGAAGAAACATCTCGGAACAACCTTTCTACTTCCCGTCAAGGCGGTTTAGCACCCGCTACGGGAACCAA
   S H V R E E T S R N N P F Y F P S R R F S T R Y G N Q
568  AACGGTAGGATCCGGTCTGACAGAGTTTGACCAAGGTCAAGGCAGTTTCAGAATCTCCAGAATCACCGTATTGTGCAG
   N G R I R V L Q R F D Q R S R Q F Q N L Q N H R I V Q
649  ATCGAGGCCAAACCTAACACTCTTGTCTTCCCAAGCAGCTGATGCTGATAACATCCTTGTATCCAGCAAGtatcaaat
   I E A K P N T L V L P K H A D A D N I L V I Q Q
730  ctaattctattctaaactacatatattttgttgcttgatacatatgattcattggattgcagGGGCAAGCCACCGTGACC
                                     G Q A T V T
811  GTAGCAATGGCAATAACAGAAAGAGCTTTAATCTTGACGAGGGCCATGCACTCAGAATCCCATCCGGTTTCTTCTAC
   V A N G N N R K S F N L D E G H A L R I P S G F I S Y
892  ATCTTGAACCGCATGACAACCAAGCTCAGAGTAGCTAAAATCTCCATGCCCGTTAACACACCCGCCAGTTTGAGta
   I L N R H D N Q N L R V A K I S M P V N T P G Q F E
973  cctctttcttttcacatatatattcaagaggaatgttaggaggccagcaacttttgaattttagttatcaaatagtcatt
1054 tagtgatgatattaatgttattaataatgattttaaaggtgtaaaattttatctaattggttcatttttttttttgataca
1135 tgttggtcagaatttaacaaattactggctcctagatttttcttatattcaattctcaattatcatcttacatgttgcttca
1216 cagGATTTCTTCCCGGCGAGCAGCCGAGACCAATCATCTACTTGCAGGGCTTCAGCAGGAATACGTTGGAGGCCGCTTC
   D F F P A S S R D Q S S Y L Q G F S R N T L E A A F
1297 AATgtaagcaaatgtatcataatttaggaattaaataacgtatcatgttataataaacttataatatatacatagCGG
   N A
1378 GAATTCAATGAGATACGGAGGGTGCTGTTAGAAGAGAATGCAGGAGGTGAGCAAGAGGAGAGAGGGCAGAGGCGATGGAGT
   E F N E I R R V L L E E N A G G E Q E E R G Q R R W S
1459 ACTCGGAGTAGTGAGAACAATGAAGGAGTGATAGTCAAAGTGTCAAAGGAGCAGCTTGAAGAACTTACTAAGCAGCTAAA
   T R S S E N N E G V I V K V S K E H V E E L T K H A K
1540 TCCGTCTCAAAGAAAGGCTCCGAAGAAGAGGAGATATCACCACCAATCAACTTGAGAGAAGGCGAGCCGCTTTCT
   S V S K K G S E E E G D I T N P I N L R E G E P D L S
1621 AACAACTTTGGGAAGTTATTTGAGGTGAAGCCAGACAAGAAGAACCCCGAGCTTCAGGACCTGGACATGATGCTCACCTGT
   N N F G K L F E V K P D K K N P Q L Q D L D M M L T C
1702 GTAGAGATCAAAGAAGGAGCTTTGATGCTCCACACTTCAACTCAAAGGCCATGTTATCGTCGTCGTAACAAAGGAAGT
   V E I K E G A L M L P H F N S K A M V I V V V N K G T
1783 GGAAACCTTGAAGTCTGCTGCTGAAGAAAGAGCAACAACAGAGGGGACGGCGGGAAGAAGAGGAGGACGAAGACGAAGAA
   G N L E L V A V R K E Q Q G R R E E E E D E E
1864 GAGGAGGGAAGTAACAGAGAGGTGCGTAGGTACACAGCGAGGTGAAGGAAGGCGATGTGTTTCATCATGCCAGCAGCTCAT
   E E G S N R E V R R Y T A R L K E G D V F I M P A A H
1945 CCAGTAGCCATCAACGCTTCTCCGAAGTCCATCTGCTTGGCTTCGGTATCAACGCTGAAAACAACCACAGAATCTTCTT
   P V A I N A S S E L H L L G F G I N A E N N H R I F L
2026 GCAGGTGATAAGGACAATGTGATAGACCAGATAGAGAAGCAAGCAAGGATTTAGCATTCCCTGGGTGGGTGAACAAGTT
   A G D K D N V I D Q I E K Q A K D L A F P G S G E Q V
2107 GAGAAGCTCATCAAAAACAGAAGGAATCTCACTTTGTGAGTGCTCGTCTCAATCTCAATCTCAATCTCCGTCGTCTCCT
   E K L I K N Q K E S H V S A R P Q S Q S Q S P S S P F
2188 GAGAAAGAGTCTCCTGAGAAAGAGGATCAAGAGGAGGAAAACCAAGGAGGGAAGGGTCCACTCCTTTCAATTTTGAAGGCT
   E K E S P E K E D Q E E E N Q G G K G P L L S I L K A
2269 TTTAACTGAGAAATGGAGGCAACTTGTTATGTATCGATAATAAGATCACGCTTTTGTACTCTACTATCCAAAACCTTATCAA
   F N *
2350 TAAATAAAACGTTTGTGCGTTGTTTCTCCAAGAAAGGTGTGGCGCTTATGGTTTTTTATTTATACGAACTAATTAATA
2431 CATCATAACGGCAAAGACCTCTTATTATGTAATTTTCTTTTAAAAATATAGACTGCCAAAATATATGAATATTTACAT
2512 GAACAAGTCAAATGGAATACTTGTGAGACATATCTCATCGATCCTATCTACTAATTAGTAGATACAAACATATTT

```

Fig. 1. Nucleotide and deduced amino acid sequences of the genomic clone of the gene encoding peanut allergen Ara h 1. In the ORF, uppercase letters indicate exon sequences and lowercase letters represent introns. The, start codon (ATG), stop codon (TGA) and polyadenylation signal (AATAAA) are underlined. The 27 amino acids putative signal peptide is italicized and underlined. The deduced amino acid sequence is below the corresponding DNA sequence.

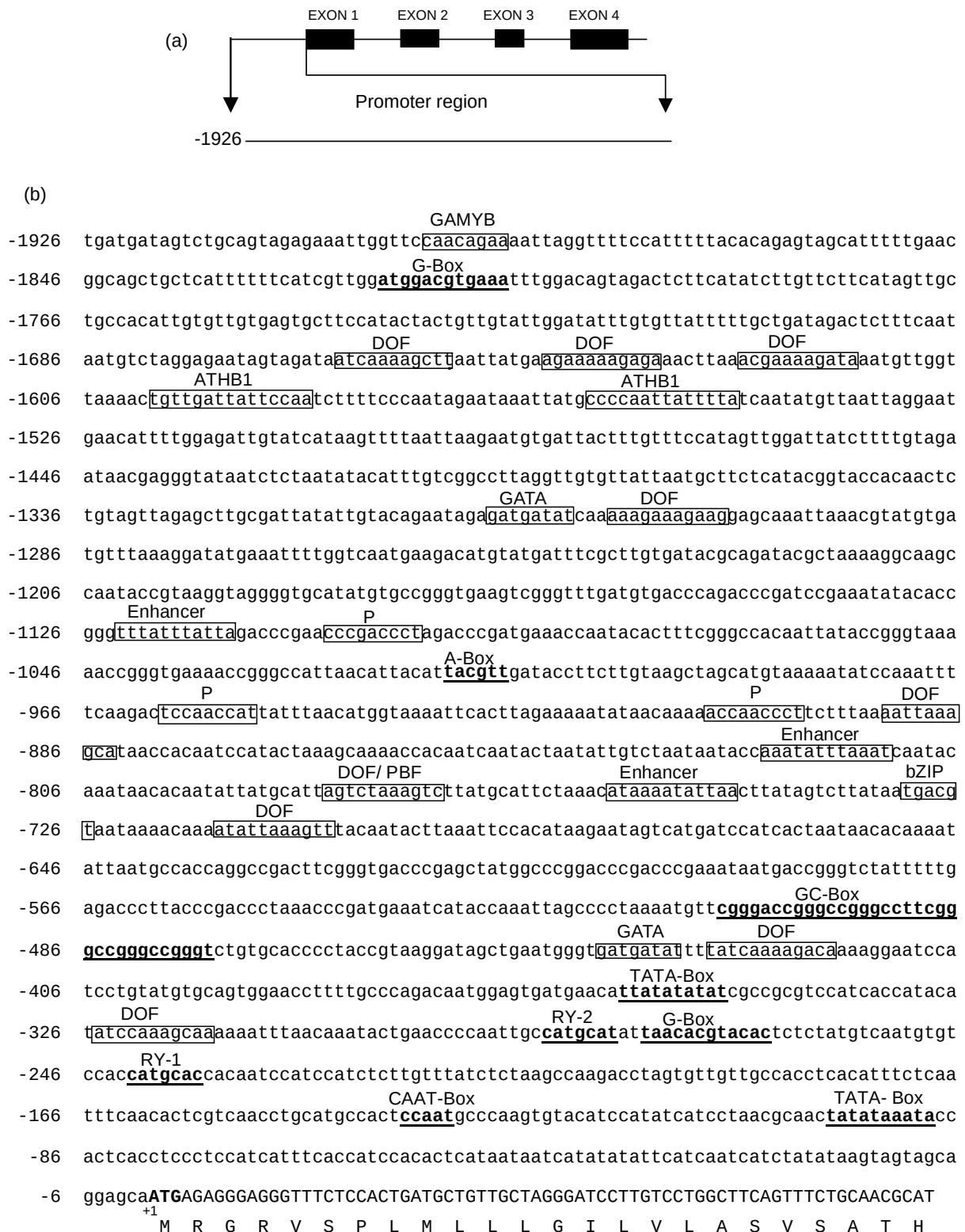


Fig. 2. Schematic diagram of the *Ara h 1* gene and sequence of the *Ara h 1* promoter. (a) The *Ara h 1* gene is composed of four exons and three introns. The promoter sequence between nucleotides -1926 and +13 is shown schematically. (b) Nucleotide sequence of the *Ara h 1* promoter region. The initiation start site (ATG) for the *Ara h 1* is bolded and shown as +1. The putative TATA-boxes, A-box, CAAT-box, GC-box, G-boxes and RY elements are bolded and underlined. Putative binding sites for transcription factors are boxed according to their position. Partial nucleotide of the first exon with its corresponding amino acid is presented.

may be considered general *Ara h 1* promoter enhancers. A motif (TGACGT) was found at position –731. A similar motif was determined in the alpha amylase gene from *Vigna mungo* (Yamauchi, 2001) and has been associated with the large family of transcription factors, such as the leucine zipper (bZIP) proteins (Tabata et al., 1989). Furthermore, two other sequences matching the RY elements or legumin boxes found in conglycinin gene (Prieto-Dapena et al., 1999), a box involved in seed-specific gene regulation (Bustos et al., 1991), were located at positions –235 (RY-1, CATGCAC) and –278 (RY-2, CATGCAT), respectively. RY elements have been related in the expression of seed's-specific genes and storage proteins (Bäumlein et al., 1992; Chamberland et al., 1992).

To date no peanut allergen promoter has been identified. Sequence analysis of this promoter revealed the presence of 17 putative DNA sequences with similarities to well-characterized *cis*-elements. The most prominent matches identified include those to the consensus binding sites for Dof proteins, for Gamyb, for Athb1, for P, and for Pbf regulatory elements (Fig. 2).

Thus, the presence of TATA, CAAT, A-, G-, GC-boxes, AT-rich regions and the identification of several transcription initiation sites suggest that the *Ara h 1* promoter is potentially a strong promoter. Further analysis will be necessary to characterize the *Ara h 1* gene promoter and to understand the underlying molecular mechanisms of the transcriptional regulation of *Ara h 1* gene expression.

3.3. The 3' flanking region of the *Ara h 1* gene

About 300 bp of the 3' UTR were sequenced. A putative polyadenylation signal (AATAAA) was located 84 nucleotides from the stop codon (ATG) (Fig. 1). This signal is highly conserved in higher eukaryotes and plays a role in positioning the poly-A tail at the end of the mRNA. Both the position and the distance separating the poly-A signal from the stop codon (TGA) are conserved between the genomic *Ara h 1* and the cDNA clones. Additional termination codons (TAA) were observed on the genomic clone at positions +2303 and +2306 or 30 and 33 bp downstream from the first termination codon (Fig. 1). Studies have shown that some plant genes have several termination codons, but it is not clear which one is preferentially used (Grierson and Covey, 1988; Berghoefer and Kloesgen, 1996). The region downstream of the poly (A) signal is T-rich and contains several TTTG/CTA/G motifs. These motifs have been associated in signaling 3' end formation in other plant poly(A) signals (Rothnie, 1996).

3.4. Amino acid sequence analysis

It was recently revealed that most food allergens are classified as either a member of the cereal prolamin superfamily or of the cupin superfamily (Breiteneder and Ebner, 2001; Dunwell et al., 2000, 2001). The cupin family is

characterized by a common β -structure thought to have originated from a prokaryotic ancestor (Dunwell et al., 2000, 2001; Khuri et al., 2001; Mills et al., 2002). Most cupins have a single domain however double domains have been identified among seed storage proteins (Dunwell et al., 2000, 2001). This structural motif is present in legumin plants such as peanut, soybean and walnut. Peanut allergen *Ara h 1* belongs to the cupin superfamily (Mills et al., 2002) and contains two domains of 143 and 123 amino acid residues. Motif 1 has 20 amino acids residues for domain 1 and 21 residues for domain 2. Motif 2 has 16 amino acids for both domains. The domains are separated by an intermotif region of 19 amino acids residues for the first domain and 45 amino acids residues for the second domain. Each of the motifs contains a single histidine.

Plant food allergens have been classified base on their amino acid in three major groups: the prolamin, cupin and cysteine protease superfamilies (Mills et al., 2003). Several seed storage proteins with known allergenicity belonging to the cupin family have been also identified as pathogenesis-related (PR) proteins (Hoffmann-Sommergruber, 2001). Recent experiments performed in our laboratory revealed that *Ara h 1* protein has trypsin inhibitory activity of 682 TIA/mg, and thus, we speculate that in addition to being a seed storage protein, *Ara h 1* may function as a PR protein and may play a role in the plant defense mechanism against pests and insects.

Acknowledgements

We thank Dr. Albert Abbott (Biological Sciences Department, Clemson University, Clemson, SC, USA) for providing the peanut genomic library and ResGen, Inc. (Huntsville, AL, USA) for sequencing.

References

- Atschul, 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.
- Bäumlein, H., Nagy, I., Villarroel, R., Inze, D., Wobus, U., 1992. Cis-analysis of a seed protein gene promoter: the conservative RY repeat CATGCATG within the legumin box is essential for tissue-specific expression of a legumin gene. *Plant J.* 2, 233–239.
- Berghoefer, J., Kloesgen, R.B., 1996. Isolation and characterization of a cDNA encoding the SecY protein from spinach chloroplasts. *Plant Physiol.* 112, 863.
- Breiteneder, H., Ebner, C., 2001. Atopic allergens of plant foods. *Curr. Opin. Allergy Clin. Immunol.* 1, 261–267.
- Burks, A.W., Williams, L.W., Helm, R.M., Connaughton, C., Cockrell, G., O'Brien, T., 1991. Identification of a major peanut allergen, *Ara hI*, in patients with atopic dermatitis and positive peanut challenges. *J. Allergy Clin. Immunol.* 88, 172–179.
- Burks, A.W., Cockrell, G., Stanley, J.S., Helm, R.M., Bannon, G.A., 1995. Recombinant peanut allergen *Ara h I* expression and IgE binding in patients with peanut hypersensitivity. *J. Clin. Invest.* 96, 1715–1721.

- Burks, W., Bannon, G.A., Sicherer, S., Sampson, H.A., 1999. Peanut-induced anaphylactic reactions. *Int. Arch. Immun.* 119, 165–172.
- Bustos, M.M., Begum, D., Kalkan, F.A., Battraw, M.J., Hall, T.C., 1991. Positive and negative *cis*-acting DNA domains are required for special and temporal regulation of gene expression by a seed storage protein promoter. *EMBO J.* 10, 1469–1479.
- Chamberland, S., Daigle, N., Bernier, F., 1992. The legumin boxes and the part of the soybean β -conglycinin promoter are involved in seed gene expression in transgenic tobacco plants. *Plant Mol. Biol.* 19, 937–949.
- Connelly, S., Marshallsay, C., Leader, D., Brown, J.W., Filipowicz, W., 1994. Small nuclear RNA genes transcribed by either RNA polymerase II or RNA polymerase III in monocot plants share three promoter elements and use a strategy to regulate gene expression different from that used by their dicot plant counterparts. *Mol. Cell. Biol.* 14 (9), 5910–5919.
- De Jong, E.C., Van Zijverden, M., Spanhaak, S., Koppelman, S.J., Pellegrom, H., Penninks, A.H., 1998. Identification and characterization of multiple major allergens in peanut proteins. *Clin. Exp. Allergy* 28 (6), 734–751.
- 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. *Micr. Mol. Biol. Rev.* 64 (1), 153–179.
- Dunwell, J.M., Culham, A., Carter, C.E., Sosa-Aguirre, C.R., Goodenough, P.W., 2001. Evolution of functional diversity in the cupin superfamily. *Trends Biochem.* 26 (12), 740–746.
- Emmett, S.E., Angus, F.J., Fry, J.S., Lee, P.N., 1999. Perceived prevalence of peanut allergy in Great Britain and its association with other atopic conditions and with peanut allergy in other household members. *Allergy* 54, 380–385.
- Foster, R., Izawa, T., Chua, N.H., 1994. Plant bZIP proteins gather at ACGT elements. *FASEB J.* 8, 192–200.
- Grierson, D., Covey, S.N., 1988. *Plant Molecular Biology*, second edition. Chapman and Hall, New York, pp. 23–46.
- Häger, K.P., Müller, B., Wind, C., Erbach, S., Fischer, H., 1996. Evolution of legumin genes: loss of ancestral introns at the beginning of the angiosperm diversification. *FEBS Lett.* 387, 94–98.
- Higgins, T.J.V., 1984. Synthesis and regulation of major proteins in seeds. *Annu. Rev. Plant Physiol.* 35, 191–195.
- Hoffmann-Sommergruber, K., 2001. Pathogenesis-related (PR)-proteins identified as allergens. *Biochem. Soc. Trans.* 30, 930–935.
- Izawa, T., Foster, R., Chua, N.H., 1993. Plant bZIP protein DNA binding specificity. *J. Mol. Biol.* 230, 1131–1144.
- Jepeal, L.I., Boylan, M.O., Wolfe, M.M., 2003. Cell-specific expression of the glucose-dependent insulinotropic polypeptide gene functions through a GATA and an ISL-1 motif in a mouse neuroendocrine tumor cell line. *Regul. Pept.* 113, 139–147.
- Kang, I.-H., Gallo-Meagher, M., 2002. Characterization of three major peanut allergen genes. *Proc. Am. Peanut Res. Educ. Soc.* 34, 97.
- Khuri, S., Bakker, F.T., Dunwell, J.M., 2001. Phylogeny, function, and evolution of the cupins, a structurally conserved, functionally diverse superfamily of proteins. *Mol. Biol. Evol.* 18, 593–605.
- Kleber-Janke, T., Crameri, 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 Appl. Immunol.* 119 (4), 265–274.
- Koppelman, S.J., Bruijnzeel-Koomen, C.A., Hessing, M., De Jongh, H.H., 1999. Heat-induced conformational changes of *Ara h1*, a major peanut allergen, do not affect its allergenic properties. *J. Biol. Chem.* 274 (8), 4770–4777.
- Law, I.J., 2000. Comparison of putative nodule and seed lectin gene promoters of peanut and in situ localization of nodule lectin gene expression. *Plant Sci.* 153 (1), 43–54.
- Lorković, Z.L., Wiczorek, D.A., Lambermon, M.H.L., Filipowicz, W., 2000. Pre-mRNA splicing in higher plants. *Trends Plant Sci.* 5 (4), 160–167.
- Maleki, S.J., Chung, S.Y., Champagne, E.T., Raufman, J.P., 2000. The effects of roasting on the allergenic properties of peanut proteins. *J. Allergy Clin. Immunol.* 106, 763–768.
- Mills, E.N.C., Jenkins, J., Marigheto, N., Belton, P.S., Gunning, A.P., Morris, V.J., 2002. Allergens of the cupin superfamily. *Biochem. Soc. Trans.* 30, 925–929.
- Mills, E.N.C., Madsen, C., Shewry, P.R., Wichers, H.J., 2003. Food allergens of plant origin—their molecular and evolutionary relationships. *Trends Food Sci. Technol.* 14, 145–156.
- 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.
- Prieto-Dapena, P., Almoguera, C., Rojas, A., Joedano, J., 1999. Seed-specific expression patterns and regulation by ABI3 of an unusual late embryogenesis-abundant gene in sunflower. *Plant Mol. Biol.* 39, 615–627.
- Qu, F., Zhai, W., Chen, H., Zhu, L.H., Moris, J., 1996. Cloning, characterization and transient expression of the gene encoding a rice U3 small nuclear RNA. *Gene* 172, 217–220.
- 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 nucleotide sequence data. *Nucl. 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 h3*. *J. Clin. Invest.* 103 (4), 535–542.
- Rothnie, H.M., 1996. Plant mRNA 3'-end formation. *Plant. Mol. Biol.* 32, 43–61.
- Sambrook, J., Fritsch, E.F., Maniatis, T., 1989. *Molecular Cloning: A Laboratory Manual*, second edition, Cold Spring Harbor Laboratory Press, New York.
- Sicherer, S.H., Munoz-Furlong, A., Burks, A.W., Sampson, H.A., 1999. Prevalence of peanut and tree nut allergy in the US determined by a random digit dial telephone survey. *J. Allergy Clin. Immunol.* 103, 559–562.
- Stanley, J.S., King, N., Burks, A.W., Huang, S.K., Sampson, H.A., Cockrell, G., Helm, R.M., West, C.M., Bannon, G.A., 1997. Identification and mutational analysis of the immunodominant IgE binding epitopes of the peanut allergen *Ara h2*. *Arch. Biochem. Biophys.* 342 (2), 244–253.
- Strömvik, M.V., Sundararaman, V.P., Vodkin, L.O., 1999. A novel promoter from soybean that is active in a complex developmental pattern with and without its proximal 650 base pairs. *Plant. Mol. Biol.* 41, 217–231.
- Tabata, T., Takase, H., Takayama, S., Mikami, K., Nakatsuka, A., Kawata, T., Nakayama, T., Iwabuchi, M., 1989. A protein that binds to a *cis*-acting element of wheat histone genes has a leucine zipper motif. *Science* 245 (4921), 965–967.
- Tatusova, T.A., Madden, T.L., 1999. Blast 2 sequences—a new tool for comparing protein and nucleotide sequences. *FEMS Microbiol. Lett.* 174, 247–250.
- 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 (4), 713–717.
- Yamauchi, D., 2001. A TGACGT motif in the 5'-upstream region of β -amylase gene from vigan mungo is a *cis*-element for expression in cotyledons of germinated seeds. *Plant. Cell Physiol.* 42, 635–641.