

Isolation of peanut genes encoding arachins and conglutins by expressed sequence tags

Yong-Sheng Yan^a, Xiao-Dong Lin^a, Yi-Shun Zhang^a,
Lei Wang^a, Keqiang Wu^b, Shang-Zhi Huang^{a,*}

^a School of Life Sciences, Zhongshan University, Guangzhou 510275, China

^b Department of Biology, West Virginia University, Morgantown, WV 26506-6057, USA

Received 2 April 2005

Available online 24 May 2005

Abstract

Seed storage proteins commonly comprise various groups of multiple isoforms encoded by different gene families. Arachin (11S globulin), conarachin (7S globulin) and conglutin (albumin) are the three major storage proteins in peanut (*Arachis hypogaea* L.). In order to isolate peanut seed storage protein genes, we sequenced and analyzed more than 400 randomly selected clones from a cDNA library derived from mid-maturation stage cotyledons of peanut. Our analysis indicated that arachin isoform precursors were encoded by at least five genes, whereas conglutins were encoded by more than six genes. The full-length cDNAs that encode isoforms of five arachin and six conglutin precursors were obtained. Among these cDNA clones, the deduced polypeptide of Ara h 1 has about 99% similarity to that of peanut allergen Ara h 3. In addition, the deduced amino acid sequences of three conglutins, AhCONG 1, AhCONG 2 and AhCONG 5, are identical to those of peanut allergens Ara h 2.02, Ara h 2 and Ara h 6, respectively. Our study identified four new members (Ara 2 to Ara 5) of arachin family and two new members (AhCONG 3 and AhCONG 6) of conglutin family in peanut.

© 2005 Elsevier Ireland Ltd. All rights reserved.

Keywords: *Arachis hypogaea*; Expressed sequence tag (EST); Arachin; Conglutin; Peanut allergen

1. Introduction

Developing seeds accumulate massive storage proteins that are used as a source of amino acids during seed germination and seedling growth. Based on their solubility in various extraction solvents [1], seed storage proteins can be classified into four groups: soluble in water (albumins), dilute saline (globulins), alcohol/water mixtures (prolamins), and dilute acid or alkali (glutelins). Storage proteins of legume are mainly composed of 11S globulins (legumin), 7S globulins (vicilin) and albumin [2,3]. Albumins are divided into two major classes, napin and conglutin. Napin 2S albumins are a class of small (1.7–2.2S) sulphur-rich proteins. In most species, napin 2S albumins are synthesized as a single polypeptide that undergoes a number of

processing steps to produce two subunits linked by disulphide bonds [4]. Napin 2S albumin exists widely in the seeds of dicotyledonous plants. Conglutins show high sequence homology to proteins of napin 2S albumin family. In contrast to napin 2S albumins, conglutins exist as a single continuous polypeptide chain in peanut seeds [5].

In peanut about 87% of the seed protein is globulin consisting of two major fractions, arachin (glycinin) and conarachin (vicilin) [6,7]. Peanuts have eminent nutritional properties, but can cause severe, type I hypersensitivity reactions. Several different peanut allergens and their genes have been isolated. The peanut allergens Ara h 1 (including Ara h 1 p41b and Ara h 1 p17), Ara h 2 (including Ara h 2.02 and Ara h 2.0), and Ara h 3/Ara h 4 could be assigned to the vicilin, conglutin, and glycinin families of seed storage proteins, respectively [8–10]. Ara h 1 and Ara h 2, both of which are major allergens, are recognized by serum IgE from over 90% of a peanut allergic patient population. In

* Corresponding author. Tel.: +86 20 84036592; fax: +86 20 84036215.
E-mail address: ls18@zsu.edu.cn (S.-Z. Huang).

contrast, Ara 3/Ara h 4, Ara h 6, and Ara h 7, which are minor allergens, are recognized by about 40–50% of the patient population. Ara h 2, and the minor allergens Ara h 6 and Ara h 7 show high sequence homology to proteins of the conglutin family from various dicotyledonous plants [9,10]. The genomic structure and organization of peanut allergens Ara h 1 p41b [11], Ara h 1 p17 (β subunit of conarachin) [12], Ara h 2.0 [13] and Ara h 3 [14] have been reported. Besides the peanut allergens, one cDNA fragment encoding an arachin isoform (accession no. AF125192) and one encoding a conglutin isoform [15] were also isolated. However, no information is available on the other members of peanut arachin and conglutin gene families.

The partial sequencing of anonymous cDNA clones (expressed sequence tags, ESTs) has become the method of choice for the rapid and cost-effective generation of data on the coding capacity of genomes. More and more genes have been isolated using ESTs. In plants, this method was initially used for the model species *Arabidopsis thaliana* [16] and *Oryza sativa* [17], and has subsequently been used in many other plant species [18–21]. Peanut EST project has been initiated recently in USA [22]. In this study, we report the identification and analysis of new members of arachin and conglutin gene families, four complete cDNAs encoding arachin and two complete cDNAs encoding conglutin precursors though EST analysis.

2. Materials and methods

2.1. RNA isolation and cDNA library construction

Total RNA was isolated from mid-maturation stage cotyledons of peanut (*Arachis hypogaea* L. Shanyou 523) using TRIZOL Reagent (GIBCOBRL, USA) according to the manufacturer's instruction. The synthesis of cDNA and an adaptor ligation were performed using the SmartTM cDNA Library Construction Kit (CLONTECH). The resulting cDNA was directly subcloned into the λ TriplEx2 vector and packaged using MaxPlaxTM lambda packaging extracts (Epicentre Technologies) according to manufacturer's guidelines. The resulting primary library contained 5×10^6 pfu.

2.2. DNA template preparation and sequencing

The conversion of a TriplEx2 phagemid to a pTriplEx2 plasmid involved in vivo excision and circularization of a complete plasmid from the recombinant phage. A total of 2 μ l phage (5 pfu/ml $\times$ 106 pfu/ml) was combined overnight with 200 μ l of BM25.8 host cell culture and the mixture was incubated at 31 °C for 30 min without shaking. LB broth (400 μ l) was added to the mixture, which was incubated at 31 °C for an additional 1 h with shaking (225 rpm). Finally, the infected cell suspension was spread on a LB/ampicillin plate to obtain isolated colonies.

Well-isolated colonies were randomly picked from each clone and plasmid DNA samples were prepared separately using the E.Z.N.A. Plasmid Minipreps DNA Purification System (Omega Bio-tek, USA). The plasmid DNA was stored at –20 °C. Purified plasmid DNA was sequenced to obtain the 5'-end sequence of the insert with a modified SMART primer (5'-CTCGGGAAGCGCGCCATTGTGTTGGT-3') and Applied Biosystems BigDye Sequencing Mix. The sequencing reactions were separated on an Applied Biosystems 3700 DNA analyzer.

2.3. Analysis of ESTs

Sequencing with the modified SMART primer produced sequence directly from the 5'-end of the insert, hence no 5'-vector sequence trimming was required. Each sequence obtained was edited using Vecscreen (<http://www.ncbi.nlm.nih.gov>) to remove flanking vector sequence and assessed manually to determine sequence quality. Comparison of the DNA sequence to non-redundant protein sequence databases at the National Centre for Biotechnology (NCBI) was performed.

2.4. Cloning of complete cDNA

Complete cDNAs of storage proteins were cloned by contig or RACE. The 5'-RACE and 3'-RACE synthetic oligonucleotide primers were constructed based on the EST sequences of storage proteins determined from the analysis of ESTs. The synthetic oligonucleotides were usual 26–30 bases in order to make sure their specificity. cDNA synthesis was performed using a SMART cDNA library construction kit (Clontech) which included an Oligo sequence (5'-AAGCAGTGGTAT-CAACGCAGAGTGGCCATTATGGC-CGGG-3') and a 5'-PCR primer (5'-AAGCA-GTGGTAT-CAACGCAGAGT-3'). The PCR products were directly subcloned into the pGEM T-Easy vector (Promega).

3. Results and discussion

3.1. Analysis of peanut storage protein ESTs

As an approach to isolate genes encoding seed proteins, over 450 random cDNA clones were sequenced and analyzed from the cDNA library which was constructed from mid-maturation stage cotyledons of peanut. Useable sequences were obtained for 414 of the clones, and seventy ESTs of arachins, conarachins and conglutins were isolated (Table 1). The abundant expressed genes were also shown in 2D PAGE (data are not shown), including *Ara 1*, *Ara 2* (encoding the isoform precursors of arachin), *Ahy α* , *Ahy β* (encoding the α and β subunits of conarachin), and *AhCONG 1*, *AhCONG 2*, *AhCONG 3* and *AhCONG 4* (encoding the isoforms of conglutin). The other arachin and conglutin genes had lower abundant expression than those mentioned above (Table 1).

Table 1
ESTs of arachins, conarachins and conglutins derived from developing peanut cotyledons

Protein	Gene name	Accession no.	Clone numbers	Rate (%)	Peanut allergen
Arachin	<i>Ara 1</i>	AY722685	8	1.93	Ara h 3
	<i>Ara 2</i>	AY722686	6	1.45	
	<i>Ara 3</i>	AY848698	2	0.48	
	<i>Ara 4</i>	AY722687	1	0.24	
	<i>Ara 5</i>	AY722688	1	0.24	
Conarachin	<i>Ahy α</i>	AY581850	3	0.72	Ara h 1 p41b
	<i>Ahy β</i>	AY581849	4	0.96	Ara h 1 p17
Conglutin	<i>AhCONG 1</i>	AY581853	9	2.17	Ara h 2.02
	<i>AhCONG 2</i>	AY722689	18	4.34	Ara h 2
	<i>AhCONG 3</i>	AY722690	9	2.17	
	<i>AhCONG 4</i>	AY848699	7	1.68	
	<i>AhCONG 5</i>	AY849314	1	0.24	Ara h 6
	<i>AhCONG 6</i>	AY722691	1	0.24	
Total ESTs			414	16.9	

Seven ESTs of conarachin were cloned which could be divided into two types, *Ahy α* and *Ahy β*. Immunoscreening of the cDNA expression library with conarachin antibody also obtained these two type cDNAs encoding conarachin isoforms in our lab [23]. The genomic structure and organization of these two genes have been reported previously [11,12].

3.2. Cloning and sequencing of arachin cDNAs

Three full-length arachin cDNA sequences named *Ara 1*, *Ara 2*, and *Ara 3* were obtained by contig. Other two full-length cDNA sequences named *Ara 4* and *Ara 5* were obtained by RACE. The *Ara 1* translated peptide has about 99% similarity to that of peanut allergen *Ara h 3* (accession no. AF093541) and about 88% similarity to that of *Ara h 4* (accession no. AF086821). Each legumin-like globulin precursor was predicted to be processed at a consensus site by an asparaginyl endopeptidase [24] and split into an acidic large subunit (A subunit) and a basic small one (B subunit) linked by disulfide bonds. The deduced amino acid sequences of arachin isoforms and their signal peptide sequences, cleavage sites are shown in Fig. 1. The putative signal amino acid residues, molecular masses and isoelectric points of the large and small subunits from the precursors of five deduced amino acid sequences of arachin isoforms are showed in Table 2. *Ara 1*, *Ara 2*, *Ara 3* and *Ara 4* are synthesized with signal peptide of 18 residues. The deduced molecular masses of the large subunits in these proteins are about 32–39 kDa, whereas all of the small subunits are about 20.5 kDa. In contrast to the other arachin isoforms, *Ara 5* is synthesized with a signal peptide of 20 residues and consists of eight conserved tetrapeptides (DDRR) in its large subunit (Fig. 1). In the mature protein, the deduced molecular masses of the two subunits of *Ara 5* are 50.3 and 23.5 kDa,

respectively (Table 2). There is no methionine in A subunit and B subunit of *Ara 3*. In comparison, there is only one methionine residue in the A subunits of the *Ara 1* and *Ara 2*, but there is no methionine in B subunits of the arachin isoforms. In contrast to *Ara 1* and *Ara 2*, there are six and five methionine residues in B subunits of the *Ara 4* and *Ara 5*, respectively, but there is no methionine residue in A subunits of the arachin isoforms.

The phylogeny tree based on the amino acid sequences of 11S globulins from different species indicates that the *Ara 1*, *Ara 2*, *Ara 3*, *Ara 4*, *Ara h 4*, and *Ara 6* (accession no. AF125192) isoforms share closer genetic relation with that of Group-I glycinin gene, whereas the *Ara 5* is evolutionary closer to that of Group-II glycinin gene in soybean seeds [25] (Fig. 3A). As predicted from the deduced amino acid sequences, the peanut arachin isoforms could be divided into three subfamilies, the first subfamily includes the *Ara 1*, *Ara 2*, *Ara 3*, *Ara 6*, and *Ara h 4* isoforms, the second and the third subfamilies contain the *Ara 4* and *Ara 5* isoforms, respectively. Percentages of sequence identity between

Table 2
Comparison of the deduced amino acid sequence of five arachin isoforms

	Arachin isoforms				
	<i>Ara 1</i>	<i>Ara 2</i>	<i>Ara 3</i>	<i>Ara 4</i>	<i>Ara 5</i>
Total residues (aa)	536	537	531	484	662
Cleavage site	SGNG	RGNG	RGNG	FNNG	GRNG
Signal peptide residues (aa)	18	18	18	18	20
MLS residues (aa) ^a	332	333	327	289	428
MLS mass (kDa)	39.0	39.2	38.1	32.1	50.3
MLS pI	5.3	5.2	5.0	5.2	5.0
MSS residues (aa) ^b	186	186	186	177	214
MSS mass (kDa)	20.6	20.4	20.7	20.6	23.5
MSS pI	6.1	6.5	7.1	5.7	8.0

^a Mature large subunits.

^b Mature small subunits.

	Signal peptide sequence	
soybean	MGKPFPTLSLSSCLLLSSACFAITSSK.FNEQQLNNLNALEPDHVESEGGIETWNSQHPELQCAGVTVSKRTLNRN	79
Ara h 4	MAK.LLELSFC.FCFLVLGASSISFRQQPEENACQFQRLNAQRPDNRIESEGGYIETWNPNNQEFECAGVALSRVLRRN	78
Ara1	MGK.LLALSVC.FCFLVLGASSISFRQQPEENACQFQRLNAQRPDNRIESEGGYIETWNPNNQEFECAGVALSRVLRRN	78
Ara2	MAK.LLALSVC.FCFLVLGASSISFRQQPEENACQFQRLNAQRPDNRIESEGGYIETWNPNNQEFECAGVALSRVLRRN	78
Ara3	MAK.LLALSVC.FCFLVLGASSISFRQQPEENACQFQRLNAQRPDNRIESEGGYIETWNPNNQEFECAGVALSRVLRRN	78
Ara4	MAK.LLALSVC.FCFLVLGASSISFRQQPEENACQFQRLNAQRPDNRIESEGGYIETWNPNNQEFECAGVALSRVLRRN	78
Ara5	MVIGPFRLSLC.VCLVFLTSACFGTRLEESFNEQQLDRNLALTPDNRIESQGGITETWNSNHPELRCAGVTLKRTIFPN	79
E1		
soybean	GSHLPSLYPQMIIIVVGKGAIGFAFGCPETFEKPOQSS.....SRRGS.RSQQLQDSHQKIRHNEGDVLV	148
Ara h 4	ALRRPFYSNAPQEIFIQQGRGYFGLIFPGCPSTYEEPAQGGRRYQSRPFRRLQEDQSQQQQ.DSHQKVRHNEGDLIA	157
Ara1	ALRRPFYSNAPQEIFIQQGRGYFGLIFPGCPSTYEEPAQGGRRYQSRPFRRLQEDQSQQQQ.DSHQKVRHNEGDLIA	157
Ara2	ALRRPFYSNAPQEIFIQQGRGYFGLIFPGCPSTYEEPAQGGRRYQSRPFRRLQEDQSQQQQ.DSHQKVRHNEGDLIA	158
Ara3	ALRRPFYSNAPQEIFIQQGRGYFGLIFPGCPSTYEEPAQGGRRYQSRPFRRLQEDQSQQQQ.DSHQKVRHNEGDLIA	157
Ara4	ALRRPFYSNAPQEIFIQQGRGYFGLIFPGCPSTYEEPAQGGRRYQSRPFRRLQEDQSQQQQ.DSHQKVRHNEGDLIA	155
Ara5	GPHLPSYANYPLIFIAQNGVFGVSLPGCPVTYEEAESQSRDRRQRIIVIKRESE.QEQEQGDSHHKIYHFGHLLA	158
E2		
soybean	IPLGVVYTYNTGDEPVVAISPLDTSFNNDQDQNPVRYLAGNPDIHEHPETMQQQ.....QQQK	208
Ara h 4	VPTGVAFWLYNDHDTDVAVSLTDNNNDNQDQFPRRNLNAGNHEQEFRLYQQQSRQSRRLSPYSPSPSRPRRER	237
Ara1	VPTGVAFWLYNDHDTDVAVSLTDNNNDNQDQFPRRNLNAGNHEQEFRLYQQQSR.....RRSLPYSPSPSPQPKQEDR	234
Ara2	VPTGVAFWLYNDHDTDVAVSLTDNNNDNQDQFPRRNLNAGNHEQEFRLYQQQSR.....RRSLPYSPSPSPQPKQEDR	235
Ara3	VPTGVAFWLYNDHDTDVAVSLTDNNNDNQDQFPRRNLNAGNHEQEFRLYQQQSRQSRRLSPYSPSPSPQPKQEDR	237
Ara4	VPHGVAFWLYNDHDTDVAVSLVHTNSLHNQDQFPRRNLNAGNHEQEFRLYQQQSG.....RQ...	214
Ara5	IPAGVPVYWSFNYGNEPIVAITLLDTSNLDNQDPSPRRYLAGNPEEHPETMQQQ.....QTRR	219
E3		
soybean	SHGGRKQ...GQHRQO.EEEGGSVLGSGSKHFLAQSFNTN.EDTAELKLRSPDDERK..QIVTVGGLS.VISPQWQE..	277
Ara h 4	EFRPRGQHSRRERAGQEEDEEGNI.FSGFTPEFLAQAFQVDDRQIVQNLNGENEEEEGAIVTVRGGLR.LLSPDTRG..	315
Ara1	EFSPRGQHSRRERAGQEEDEEGNI.FSGFTPEFLAQAFQVDDRQIVQNLNGENEEEEGAIVTVRGGLR.LLSPDRKRRQ	313
Ara2	EFSPRGQHSRRERAGQEEDEEGNI.FSGFTPEFLAQAFQVDDRQIVQNLNGENEEEEGAIVTVRGGLR.LLSPDRKRRQ	314
Ara3	EFSPRGQHSRRERAGQEEDEEGNI.FSGFTPEFLAQAFQVDDRQIVQNLNGENEEEEGAIVTVRGGLR.LLSPDRKRG.	315
Ara4	..SPKG.....EEQEQEENEGGNI.FSGESTPEFLSHGFQVN.EDIVRNLRGENEEEEGAIVTVRGGLS.LLVP.....	279
Ara5	RHGQHQQ...DEYGSQEEEGNNVLSGFESTQLLAHAFGVN.EEIAIRILQNPPEQTKD.QIVRVEGGFRDVISPRWEGEK	293
E4		
soybean	QEDEDE.....DEDEEYGRTPS	294
Ara h 4	ADEEE.....	320
Ara1	QYE.....RPDEEE.....	322
Ara2	QYE.....RPDEEE.....	323
Ara3	ADEEE.....	320
Ara4	PEWRO.....	284
Ara5	QYDELEERQRPDRDEQGGYDYDDRRPRHRQDPYREGDEDDRRPRGSRQGQGRGYDDDDRRPGQYEEGEEDDRRPR	373
A subunit B subunit		
soybean	YPPRRPSHG.KHEDDE...DEDE...EEDQPRPDHPPQR....PSRPEQGEPRGRG.....CQTRNGVEE	348
Ara h 4	EYDEDEQ.....Y EYHEQ.....DGRR.....GRGSRGGG.....NGIEE	349
Ara1	EYDEDEQ.....Y EYDEEERQHDRR.....GRGSRGGG.....NGIEE	355
Ara2	EYDEDEQ.....Y EYDEEERQHDRR.....GRGSRGGG.....NGIEE	356
Ara3	EYDEDEQ.....Y EYDEE.....DRRR.....GRGSRGGG.....NGIEE	349
Ara4	SYQP.....GRGDKDFN.....NGIEE	302
Ara5	RSSRPRKQRRHDDDRRADEDDRRGYDDDERRDEDDRRGYDDDDRRPRWSSRPKGQGRNGVEE	453
E4		
soybean	NICTMKLHENIARPSRADFYNPAGKAGRISTLNSLT.L.PAL.RQFGLSAQYVVLRYNGIYSPDWNLNANSVTM.TRGKGRVR	427
Ara h 4	TICTACVKKNIGNRSPDIYNPQAGSLKTANLNL.L.LL.RWLGLSAEYGNLYRNALFVPHYNTNAHSIIYALRGRAHVQ	429
Ara1	TICTASFKKNIGNRSPDIYNPQAGSLKTANLNL.L.LL.RWLGLSAEYGNLYRNALFVPHYNTNAHSIIYALRGRAHVQ	435
Ara2	TICTASVKKNIGNRSPDIYNPQAGSLKTANLNL.L.LL.RWLGLSAEYGNLYRNALFVPHYNTNAHSIIYALRGRAHVQ	436
Ara3	TICTASVKKNIGNRSPDIYNPQAGSLKTANLNL.L.LL.RWLGLSAEYGNLYRNALFVPHYNTNAHSIIYALRGRAHVQ	429
Ara4	TICTATVKKNIGKSTADIIYNPQAGSVRTVNELDL.PIL.NRLGLSAEYGSIRHDMFVPHYNNMANSIMYALHGGAHVQ	382
Ara5	TLCSPTLVEDIARPSRADFYNPAAGRISSANSLTF.PIL.RWFQLSAEHVLLRYNGIYSPHWNANNSIIYGLRGEGRIQ	533
E4		
soybean	VVNCQGNVAFDGLRRGQLLVVPQNPFAVAGQSGEGLVYVFKTHNAVSS.....YIKDVR...	483
Ara h 4	VVDSNGNRVYDEELQEGHVLVVPQNPFAVAGKSQSENFEYVAFKTDSPRSIANFAGENSFIDN.....	489
Ara1	VVDSNGDRVFDEELQEGHVLVVPQNPFAVAGKSQSENFEYVAFKTDSPRSIANLAGENSFIDN.....	495
Ara2	VVDSNGNRVYDEELQEGHVLVVPQNPFAVAGKSQSDNFEYVAFKTDSPRSIANLAGENSIIIDN.....	496
Ara3	VVDSNGNRVYDEELQEGHVLVVPQNPFAVAGKSQSDNFEYVAFKTDSPRSIANLAGENSVIDN.....	489
Ara4	VVDCNGNRVYDEELQEGHVLVVPQNPFAVAGKSQSEHFLYVAFKTDSPRSIANLAGKNSYMN.....	442
Ara5	VVNSQGNVAFVGLREGQILLVVPQNPFAVAGKQAGNEGFYVAFKTDSPRSIANLAGKNSYMN.....	611
E4		
soybean	VIPSEVLNSNSYLGQSQRQLKYQGNSGP.LVNP.....	516
Ara h 4	LPEEVVANSYGLPREQAR.QLKNNNPFKFFVPPQ.QSPRAVA	530
Ara1	LPEEVVANSYGLPREQAR.QLKNNNPFKFFVPPSE.QSLRAVA	536
Ara2	LPEEVVANSYGLPREQAR.QLKNNNPFKFFVPPSQ.QSLGAVA	537
Ara3	LPEEVVANSYGLPREQAR.QLKNNNPFKFFVPPSQ.QSPRAVA	531
Ara4	LPEDVVANSYGLQYEQAR.QLKNNNPFVFLVPPQDSQMIRTV	484
Ara5	TMETPLWSLLMI.PSMGLNVVINLNNNNAQVDSKNNDSRLWPPSS.IIHK...	662

Fig. 1. Sequence alignment of precursors of a soybean glycinin (accession no. P04347), peanut allergen Ara h 4 (accession no. AF082861) and arachin isoforms. The amino acid number for the last residue in each line is listed on the right for each arachin. Broken lines in the sequences represent gaps introduced for best alignment. The cleavage site of the putative N-terminal signal sequence and the consensus asparaginyl cleavage site that splits these aligned polypeptides into A (acidic) and B (basic) subunits are indicated by scissor symbols individually. Four conserved cysteine residues, DDDR repeats and four IgE-binding epitopes (E1–E4) are shaded. The epitope amino acids that led to decrease in IgE binding when altered [26] are shown as the bold and underlined residues.

Signal Peptide sequence			✂
AhCONG1	MAKLTI ~~IL~~ VALA.LFLAAHASA.RQWELQG...DRR.....CQS ~~QL~~ ERANL.RPCEQHLMQKIQRDEDSYER <u>DPYSP</u> SQ	69	
AhCONG2	MAKLTI ~~IL~~ VALA.LFLAAHASA.RQWELQG...DRR.....CQS ~~QL~~ ERANL.RPCEQHLMQKIQRDEDSYER <u>DPYSP</u> SQ	69	
AhCONG3	MAKSTI ~~IL~~ VALL.ALVLVAHASAMRRERGRQG...DSSS.....CERQVDRVNL.KPCEQHIMQIRIMGEQEYDS.....	64	
AhCONG4	MAKSTI ~~IL~~ VALL.ALVLVAHASAMRRERGRQG...DSSS.....CERQVDRVNL.KPCEQHIMQIRIMGEQEYDS.....	64	
AhCONG5	MAKSTI ~~IL~~ VALL.ALVLVAHASAMRRERGRQG...DSSS.....CERQVDRVNL.KPCEQHIMQIRIMGEQEYDS.....	48	
AhCONG6	MVKLSI ~~IL~~ VALLGALLVVASATRWDPDRGSRGLRWDAPS ~~RG~~ DDCQ ~~Q~~ RQLQRANL.RPCEEHIRQ ~~RV~~ EQE ~~Q~~ EQEY <u>PY</u> SQR	80	
Ara h 7	MVKLSI ~~IL~~ VALLGALLVVASATRWDPDRGSRGSRWDAPS ~~RG~~ DDCQ ~~Q~~ RQLQRANL.RPCEEHMR ~~RR~~ VEQE ~~Q~~ EQEY <u>PY</u> SRR	80	
			E1 E2
AhCONG1	<u>DPYSP</u> SQDPDR <u>DPYSP</u> SPYDRRGAGSSQH ~~Q~~ ERCNELNEFENNQRCMCEALQQIMENQSDRLQ.GRQEQEQFKRELRLNL	148	
AhCONG2	<u>DPYSP</u> S.....PYDRRGAGSSQH ~~Q~~ ERCNELNEFENNQRCMCEALQQIMENQSDRLQ.GRQEQEQFKRELRLNL	136	
AhCONG3	YDIRSTRSSDQQRCCDELNEMENTQRCMCEALQIMENQCDRLQ.DRQMVQEQFKRELRLNL	124	
AhCONG4	YDIRSTRSSDQQRCCDELNEMENTQRCMCEALQIMENQCDRLQ.DRQMVQEQFKRELRLNL	124	
AhCONG5	YNFGSTRSSDQQRCCDELNEMENTQRCMCEALQIMENQCDGLQ.DRQMVQHFRELRLNL	108	
AhCONG6	G.....SRGRPGESDEDEQ ~~Q~~ RCNELNRFQNNQRCMCQALQ ~~Q~~ ILQNSFRFQ ~~Q~~ DRS ~~Q~~ LHQNGEGAQEL	144	
Ara h 7	G.....SRGRPGESDENQ ~~Q~~ RCNELNRFQNNQRCMCQALQ ~~Q~~ ILQNSFWVPAGQEPVADSGEGAQEL	144	
			E3 E4
AhCONG1	PQQCGLRAPQRC ~~DL~~ EVESGGRDRY	172	
AhCONG2	PQQCGLRAPQRC ~~DL~~ DVESGGRDRY	160	
AhCONG3	PQQCNFRAPQRC ~~DL~~ DV.SGGRC..	145	
AhCONG4	PQQCNFRAPQRC ~~DL~~ DV.SGGRC..	145	
AhCONG5	PQQCNFRAPQRC ~~DL~~ DV.SGGRC..	145	
AhCONG6	APELRVQVTKPLRP.....	158	
Ara h 7	APELRVQVTKPLRPL.....	159	

Fig. 2. Sequence alignment of conglutin precursors in peanut seeds. The amino acid number for the last residue in each line is listed on the right for each species. Broken lines in the sequences represent gaps introduced for best alignment. The cleavage site of the putative N-terminal signal sequence is indicated by scissor symbols. Conserved cysteine residues, which probably form intrachain disulfide bonds are shaded. Three IgE-binding epitopes (E1–4) are shaded. The epitope amino acids that led to a decrease in IgE binding when altered, [9] are shown as the bold and underlined residues.

members in first subfamily range from 88–96%, but percent identities between members of different groups are less than 70%.

Peanut allergy is a serious IgE-mediated health problem because of the increased prevalence, potential severity, and chronicity of the reaction. Rabjohn et al. [26] have found four IgE-binding epitopes between 10 and 15 amino acids in Ara h 3 primary sequence. These epitope motifs are designated as epitope 1 (IETWN**PNN**QEFECAG), epitope 2 (GN**IFSGFT**-PE**FLAQA**), epitope 3 (VTVRGG**LRI**LSPDR), and epitope 4 (DEDEY**EYDEEDRG**) [26]. Epitope 3 was recognized by serum IgE from 100% of the Ara h 3-allergic population, whereas the other epitopes were recognized by serum IgE from 25 to 38% of the Ara h 3-allergic population. Mutation analysis of the epitopes revealed that single or critical amino acid changes within these peptides could lead to a reduction or loss of IgE binding [26]. As shown in Fig. 1, Ara h 4, Ara 1, Ara 2, and Ara 3 have nearly identical four epitopes, whereas Ara 4 has only three epitopes (epitope 1, 2, and 3). Besides Ara h 4 and Ara h 3 (Ara 1), Ara 2, Ara 3, and Ara 4 could be the new members of peanut allergens. Compared with other isoforms of arachin, Ara 5 also has three similar epitope motifs (epitope 1–3), but some critical amino acid residues are mutated in each motifs. These mutations may lead to complete loss or decrease in IgE binding. Further analysis is required to determine whether Ara 5 is an allergen. If the epitopes of Ara 5 have reduced or lost IgE binding, it is possible to alter the Ara h 3- or Ara h 4-like allergen with a reduced IgE-binding capacity by mutagenesis.

3.3. Cloning and sequencing of conglutin cDNAs

In most species, napin 2S albumin is synthesized as single polypeptide precursor and splits into two subunits, which are linked by disulfide bonds after processing. In contrast to 2S

albumins, peanut conglutins including Ara h 2, Ara h 6 and Ara h 7, exist as a single continuous polypeptide chain in seeds [5,10]. Six full-length cDNA clones encoding conglutin precursors, AhCONG 1 to AhCONG 6, were sequenced in our present study. Fig. 2 shows the deduced amino acid sequences of conglutin isoforms. AhCONG 1 to AhCONG 5 has eight conserved cysteine residues in their translated peptides and is synthesized with signal peptide of 20 residues. In comparison, AhCONG 6 has six conserved cysteine residues in the translated peptide and is synthesized with signal peptide of 21 residues similar to the peanut allergen Ara h 7. AhCONG 1, AhCONG 2 and AhCONG 5 translated peptides have 100% identity to the peanut allergen Ara h 2.02, Ara h 2.0 and Ara h 6. AhCONG 3 translated peptides has 98% identity to that of AhCONG 4, whereas AhCONG 6 translated peptides has 70% identity to Ara h 7. The putative signal amino acid residue numbers, molecular masses and isoelectric points of the mature proteins from six deduced amino acid sequences of conglutin precursors are showed in Table 3. While the deduced AhCONG 3, AhCONG 4 and AhCONG 5 mature proteins contained

Table 3
Comparison of the deduced amino acid sequence of six conglutins

	AhCONG (conglutin isoforms)					
	1	2	3	4	5	6
Total residues	172	160	145	145	145	158
Signal peptide residues (aa)	21	21	21	21	21	20
MP residues (aa) ^a	151	139	124	124	124	138
MP mass (kDa)	18.0	16.6	14.8	14.8	14.5	16.5
MP pI	5.5	5.5	5.5	5.3	5.0	6.1
Number of cysteine in MP	8	8	10	10	10	6
Percentage of cysteine (%)	5.3	5.7	8.1	8.1	8.1	4.3
Number of methionine in MP	3	3	8	8	8	1
Percentage of methionine (%)	1.9	2.2	6.5	6.5	6.5	0.8

^a Mature protein.

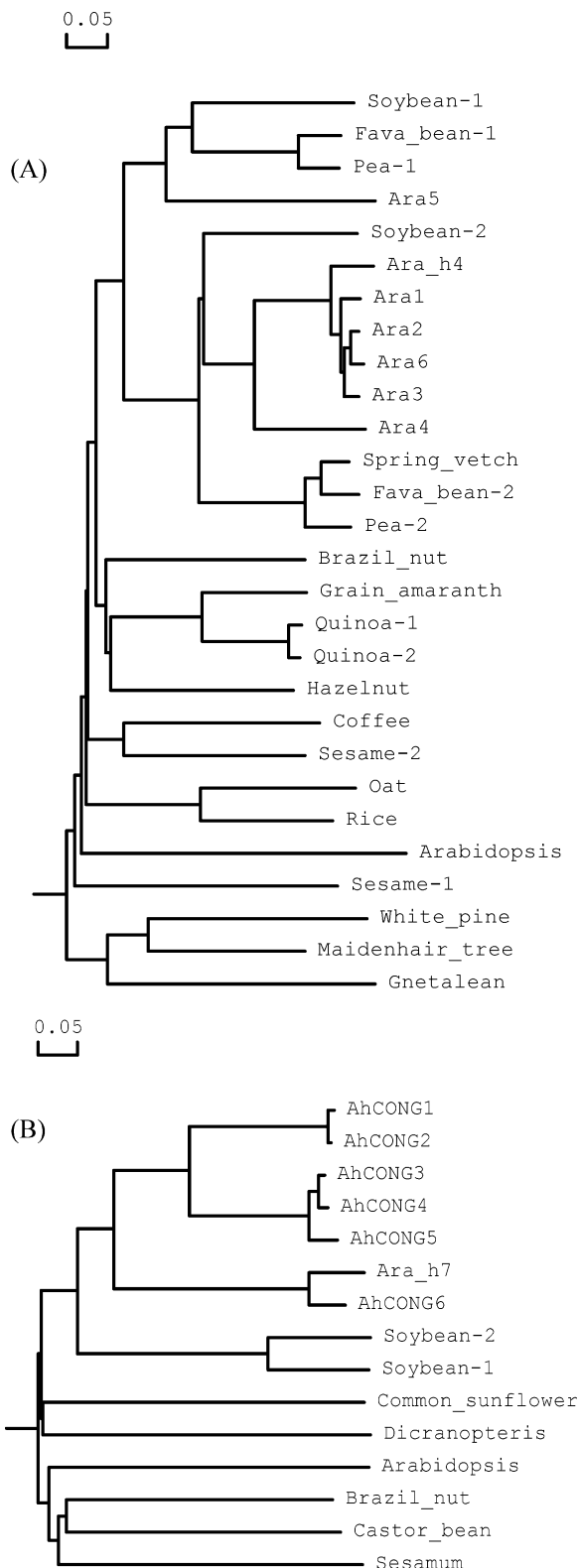


Fig. 3. Phylogeny trees. The phylogeny trees were built up with MegAlign program from DNAMAN software based on the amino acid sequences of storage proteins from different species. (A) The *Arachis hypogaea* arachin isoforms were compared with the 11S globulins of *Sesamum indicum*-1 (AF091842), *Sesamum indicum*-2 (AF240004), *Chenopodium quinoa*-1 (AY562550), *Chenopodium quinoa*-2 (AY562549), *Vicia faba*-1 (X03677), *Vicia faba*-2 (X55013), *Arabidopsis thaliana* (NM179134),

the highest amounts (6.5%) of methionine, the methionine contents of deduced AhCONG 1 and AhCONG 2 mature proteins were 1.9 and 2.2%, respectively. The methionine content of deduced AhCONG 6 mature proteins was only 0.8%. Recently, Paik-Ro et al. [15] had isolated the cDNA fragment *Psc 33* whose translated peptide has about 75% similarity to that of *Ara h 2.0* and about 94% similarity to that of *Ara h 6*. In contrast, *Psc 33* has 100% sequence identity to *AhCONG 4*.

The phylogeny based on the amino acid sequence from peanut conglutins and from the napin 2S albumins of different species indicates that the peanut conglutin isoforms could be divided into three classes, class I (AhCONG 1 and AhCONG 2), class II (AhCONG 3), and class III (AhCONG 6 and Ara h 7). There is 90% identity within the conglutin isoforms of the class, and there is 35–5.5% identity between isoforms of the different classes. Nearly all *Arachis* species are diploid, but cultivated peanut is an allotetraploid (genome AABB). It has been suggested that *A. villosa* (A genome) and *A. ipaënsis* (B genome) are the progenitors of cultivated peanut [27]. Since peanut Shanyou 523 is an outbreed line, it is possible that the pairs of similar sequences (AhCONG 1 + 2, AhCONG 3 + 4 + 5) are allelic. It remains to be determined if these similar sequences were alleles of the same gene, or two very similar genes present in different genomic locations. The phylogeny tree also indicates that the peanut conglutins are evolutionary closer to napin 2S albumins in soybean seeds (Fig. 3B). We choose Shanyou 523 for analysis because it is economic important and widely used variety in China.

Stanley et al. [9] have found three IgE-binding epitopes with 10 amino acids in Ara h 2 primary sequence (the same as AhCONG 2). These epitope motifs are epitope 1 (DRRCQSQLER), epitope 2 (YERDPYSPSQ), and epitope 3 (SQDPYSPSPY), as shown in Fig. 2. These epitopes were recognized by all patients tested. Mutation analysis of the Ara h 2 epitopes indicated that single amino acid changes result in loss of IgE binding. Two epitopes (E2 and E3) that contain the amino acid sequence **DPYSP** are necessary for IgE binding [9]. Besides epitopes 1, 2 and 3, Ara h 2.02 (AhCONG 1) has the fourth epitope (RRDPYSPSPY) that also contains the critical amino acid sequence **DPYSP**. Although there is no obvious IgE-binding epitope in AhCONG 6 and Ara h 7, the critical amino acid sequence **PYS** exists in the region between amino acids 76–8 of these

Bertholletia excelsa (AY221641), *Corylus avellana* (AF449424), *Welwitschia mirabilis* (Z50780), *Coffea Arabica* (AF054895), *Amaranthus hypochondriacus* (X82121), *Pinus strobes* (Z11487), *Ginkgo biloba* (Z50778), *Vicia sativa* (Z32835), *Avena sativa* (X74740), *O. sativa* (X54313), *Pisum sativum*-1 (X07014), *Pisum sativum*-2 (X02982), *Glycine max*-1 (M10962), *Glycine max*-2 (X15122), and compared with peanut allergen Ara h 4 (AF082861). (B) The *A. hypogaea* conglutins were compared with the napin 2S albumins of *Arabidopsis thaliana* (P15460), *Glycine max*-1 (P19594), *Glycine max*-2 (AAD09630), *Ricinus communis* (P01089), *Bertholletia excelsa* (P04403), *Helianthus annuus* (P23110), *Dicranopteris curranii* (AAF17602), *Sesamum indicum* (AF09184), and compared with peanut allergen Ara h 7 (AAD56719).

two minor allergens. AhCONG 3, 4, and 5 belong to the group of peanut allergen Ara h 6. No linear IgE reactive epitopes were identified in AhCONG 3, 4, and 5. The locations of these linear epitopes in the three-dimensional structure were addressed by building a molecular model of Ara h 2 on the basis of the experimentally solved structure of Ara h 6. The mutations in AhCONG 3, 4, and 5 proteins could lead to a reduction in IgE binding. It might be possible to create mutations in the major allergens, Ara h 2 and Ara h 2.02, by mutations of epitope 1 and deletion of epitopes 2, 3, so that they cannot bind to IgE, and 4. This raises the possibility that altered allergen genes could be used to replace their allergenic homologues in the peanut genome to produce a hypoallergenic peanut.

Acknowledgments

The work was supported by Natural Science Foundation of Guangdong Province (021695) and by National Natural Science Foundation of China (30370912). We appreciate comments and suggestions made by two anonymous reviewers.

References

- [1] P.R. Shewry, Plant storage protein, *Biol. Rev.* 70 (1995) 375–426.
- [2] A. Vitale, R. Bollini, Legume storage proteins, in: J. Kigel, G. Galili (Eds.), *Seed Development and Germination*, Marcel Dekker, New York, USA, 1995, pp. 73–102.
- [3] N. Lambert, J.N. Yarwood, Engineering legume seed storage proteins, in: P.R. Shewry, S. Gutteridge (Eds.), *Plant Protein Engineering*, Cambridge University Press, London, UK, 1992, pp. 167–187.
- [4] M.L. Ericson, J. Rodin, M. Lenman, K. Glimelius, L.G. Josefsson, L. Rask, Structure of the rape seed 1.7S storage protein, napin, and its precursor, *J. Biol. Chem.* 261 (1986) 14576–14581.
- [5] K. Lehmann, S. Hoffmann, P. Neudecker, M. Suhr, W.M. Becker, P. Rosch, High-yield expression in *Escherichia coli*, purification, and characterization of properly folded major peanut allergen Ara h 2, *Protein Exp. Purif.* 31 (2003) 250–259.
- [6] S.M.M. Basha, S.K. Pancholy, Changes in the polypeptide composition of maturing seeds from four peanut (*Arachis hypogaea* L.) cultivars, *Peanut Sci.* 8 (1981) 6–10.
- [7] J. Mosse, J.C. Pernollet, Storage protein of legume seeds, in: S.K. Arora (Ed.), *Chemistry and Biochemistry of Legumes*, Edward Arnold, Delhi, 1983, pp. 111–193.
- [8] A.W. Burks, G. Cockrell, J.S. Stanley, R.M. Helm, G.A. Bannon, Recombinant peanut allergen Ara h 1 expression and IgE binding in patients with peanut hypersensitivity, *J. Clin. Invest.* 96 (1995) 1715–1721.
- [9] J.S. Stanley, N. King, A.W. Burks, S.K. Huang, H. Sampson, G. Cockrell, R.M. Helm, C.M. West, G.A. Bannon, Identification and mutational analysis of the immunodominant IgE binding epitopes of the peanut allergen Ara h 2, *Arch. Biochem. Biophys.* 342 (1997) 244–253.
- [10] T. Kleber-Janke, R. Crameri, U. Appenzeller, M. Schlaak, W.M. Becker, Selective cloning of peanut allergens, including profiling and 2S albumins, by phage display technology, *Int. Arch. Allergy Appl. Immunol.* 119 (1999) 265–274.
- [11] O.M. Viquez, K.N. Konan, H.W. Dodo, Structure and organization of the genomic clone of a major peanut allergen gene, Ara h 1, *Mol. Immunol.* 40 (2003) 565–571.
- [12] H.G. Li, L. Wang, Y.S. Zhang, B. Liao, Y.S. Yan, X.D. Lin, S.Z. Huang, Cloning and sequencing of the gene Ah γ - β encoding a subunit of peanut conarachin, *Plant Sci.* 168 (2005) 1387–1392.
- [13] O.M. Viquez, C.G. Summer, H.W. Dodo, Isolation and molecular characterization of the first genomic clone of a major peanut allergen, Ara h 2, *J. Allergy Clin. Immunol.* 107 (2001) 713–717.
- [14] O.M. Viquez, K.N. Konan, H.W. Dodo, Genomic organization of peanut allergen gene, Ara h 3, *Mol. Immunol.* 41 (2004) 1235–1240.
- [15] O.G. Paik-Ro, J.C. Seib, R.I. Smith, Seed-specific, developmentally regulated genes of peanut, *Theor. Appl. Genet.* 104 (2002) 236–240.
- [16] S.D. Rounsley, A. Glodek, G. Sutton, M.D. Adams, C.R. Somerville, J.C. Venter, A.R. Kerlavage, The construction of *Arabidopsis* expressed sequence tag assemblies/a new resource to facilitate gene identification, *Plant Physiol.* 112 (1996) 1177–1183.
- [17] K. Yamamoto, T. Sasaki, Large-scale EST sequencing in rice, *Plant Mol. Biol.* 35 (1997) 135–144.
- [18] F. Sterky, S. Regan, J. Karlsson, M. Hertzberg, A. Rohde, A. Holmberg, B. Amini, R. Bhalerao, M. Larsson, R. Villarroel, M. Van Montagu, G. Sandberg, O. Olsson, T.T. Teeri, W. Boerjan, P. Gustafsson, M. Uhlen, B. Sundberg, J. Lundberg, Gene discovery in the wood-forming tissues of poplar/analysis of 5692 expressed sequence tags, *Proc. Natl. Acad. Sci. USA* 95 (1998) 13330–13335.
- [19] E. Asamizu, Y. Nakamura, S. Sato, S. Tabata, Generation of 7137 non-redundant expressed sequence tags from a legume, *Lotus japonicus*, *DNA Res.* 7 (2000) 127–130.
- [20] R. Shoemaker, P. Keim, L. Vodkin, E. Retzel, S.W. Clifton, R. Waterston, D. Smoller, V. Coryell, A. Khanna, J. Erpelding, X.W. Gai, V. Brendel, C. Raph-Schmidt, E.G. Shoop, C.J. Vielweber, M. Schmatz, D. Pape, Y. Bowers, B. Theising, J. Martin, M. Dante, T. Wylie, C. Granger, A compilation of soybean ESTs: generation and analysis, *Genome* 45 (2001) 329–338.
- [21] J. Fernandes, V. Brendel, X. Gai, S. Lal, V.L. Chandler, R.P. Elumalai, D.W. Galbraith, E.A. Pierson, V. Walbot, Comparison of RNA expression profiles based on maize expressed sequence tag frequency analysis and micro-array hybridization, *Plant Physiol.* 128 (2002) 896–910.
- [22] M. Luo, P. Dang, B. Guo, C. Holbrook, M. Bauscher, R. Lee, Generation of expressed sequence tags (EST) for gene discovery and marker development in cultivated peanut, *Crop Sci.* 45 (2005) 346–353.
- [23] L. Wang, Y.S. Yan, B. Liao, X.D. Lin, S.Z. Huang, The cDNA cloning of conarachin and the gene expression in developing peanut seeds, *J. Plant Physiol. Mol. Biol.* 31 (2005) 107–110.
- [24] E.C. De Jong, M. Van Zijverden, S. Spanhaak, S.J. Koppelman, H. Pellegrom, A.H. Penninks, Identification and characterization of multiple major allergens in peanut proteins, *Clin. Exp. Allergy* 28 (1998) 734–751.
- [25] N.C. Nielsen, C.D. Dickinson, T. Cho, V.H. Thanh, B.J. Scallan, R.L. Fischer, T.L. Sims, G.N. Drews, R.B. Goldberg, Characterization of the glycinin gene family in soybean, *Plant Cell* 1 (1989) 313–328.
- [26] P. Rabjohn, E.M. Helm, J.S. Stanley, C.M. West, H.A. Sampson, A.W. Burks, G.A. Bannon, Molecular cloning and epitope analysis of peanut allergen Ara h 3, *J. Clin. Invest.* 103 (1999) 535–542.
- [27] M.C. Moretzsohn, M.S. Hopkins, S.E. Mitchell, S. Kresovich, J.F. Valls, M.E. Ferreira, Genetic diversity of peanut (*Arachis hypogaea* L.) and its wild relatives based on the analysis of hypervariable regions of the genome, *BMC Plant Biol.* 4 (2004) 11.