Biochem Genet (2012) 50:372–388
DOI 10.1007/s10528-011-9481-1

Vicilin Genes of *Vigna luteola*: Structure, Organization, Expression, and Variation

Zhongyu Xie · Joseph Neigel · Caryl Chlan

Received: 2 April 2011 / Accepted: 6 November 2011 / Published online: 9 December 2011

Abstract Two different but related sequences that encode *Vigna luteola* 7S vicilins were isolated and characterized. The sequences differ by two nucleotide substitutions, each of which results in an amino acid replacement. This low level of divergence suggests that a recent gene duplication has occurred. Both variants are expressed in cDNA populations; therefore, neither gene is a pseudogene. Both copies were present in all individuals (72) analyzed using real-time PCR and TaqMan probes. Segregation was not observed. The two sequences are not independent alleles. Vicilin genomic sequences of 11 specimens from six geographic locations were determined. No polymorphic sites were identified in either of the two gene copies. This lack of polymorphism suggests that either a population bottleneck or selection has occurred. The genetic structure, expression patterns, and protein composition of the *V. luteola* vicilins were compared to those of other legume vicilins.

Keywords Vicilin · Storage protein · Gene duplication · Polymorphism · *Vigna luteola*

Introduction

Seed storage proteins provide nitrogen and amino acids for germinating seeds and seedlings. In mature legume seeds, the seed storage proteins may contribute as much

Z. Xie · J. Neigel · C. Chlan (✉)
Biology Department, University of Louisiana at Lafayette,
P.O. Box 42451, Lafayette, LA 70504, USA
e-mail: cchlan@louisiana.edu

Present Address:
Z. Xie
Ben May Department for Cancer Research, University of Chicago,
929 E. 57th Street, Chicago, IL 60637, USA

as 80% of seed nitrogen that accumulates during seed development (Pernollet and Mossé 1983). Globulins (soluble in dilute salt solutions), albumins (soluble in water), and prolamins (soluble in alcohol) are the major seed storage protein groups (Osborne 1924). The salt soluble globulins are further divided into vicilins and legumins. Vicilins form 7S aggregates and are usually glycosylated; legumins form 11S aggregates and are not usually glycosylated (Shewry and Halford 2002). Legumins and vicilins both accumulate to high levels during embryogenesis and are stored in cotyledon tissues of developing legume seeds.

Previous studies of vicilin genes and their protein products have shown a high degree of divergence at the nucleotide and amino acid sequence levels (Lawrence et al. 1994). At first, such divergence was not considered surprising, because the major function of these proteins was to supply nitrogen and amino acids for germination and early plant development. Crystallographic studies and subsequent comparisons of predicted protein structures of vicilins, however, have shown that the three-dimensional structures of vicilins are highly conserved (Ko et al. 1993, 2000; Lawrence et al. 1994).

To examine how vicilin genes have evolved with such a high degree of divergence at the primary sequence level, yet have retained a conserved hierarchical structure, we decided to look at the structure, organization, and relationships of vicilin genes in wild legume populations. Most of our knowledge of vicilin genes and proteins is from studies of nutritionally important crops such as *Glycine max* (Harada et al. 1989), *Phaseolus vulgaris* (Slightom et al. 1983), *Canavalia ensiformis* (Ng et al. 1993), *Pisum sativum* (Bown et al. 1988), and *Arachis hypogaea* (Viques et al. 2003). Since, many of these agricultural cultivars are highly inbred, they are not optimal candidates for the study of the roles that natural selection or genetic constraints might play in the conservation of hierarchical structures with highly variable primary sequences. To address these questions, we decided to focus first on characterization of the vicilin sequences within and between individuals in populations of a local wild legume species. *Vigna luteola* was chosen for this study because it occurs in multiple locations in Louisiana, has seeds that are relatively large and easy to extract, is easily identified, and large tracts of geographically isolated populations are located within reasonable distances. In addition, vicilin sequences from three related species, *V. angularis*, *V. radiata,* and *V. unguiculata,* were available for comparative analyses.

Materials and Methods

Plant Materials

Fresh leaf tissue from 75 individuals of *V. luteola* was collected from nine locations ranging across 70 miles of southern Louisiana. All samples were stored at −80°C before genomic DNA extraction. GPS coordinates were recorded for samples from all of the collection sites, and pressed reference specimens are archived in the Biology Department Herbarium at the University of Louisiana, Lafayette.

Isolation of Genomic and Plasmid DNA

Small-scale genomic DNA preparations were obtained using the PureGene DNA Purification Kit (Gentra Systems), following the manufacturer's directions. Large-scale preparations of *V. luteola* genomic DNA were purified as previously described (Li et al. 2001; Paterson et al. 1993). Plasmid DNAs were purified from overnight bacterial cultures using the QiaPrep system, as recommended by the manufacturer (Qiagen).

Amplification and Cloning of Vicilin Genomic Sequences

The derived amino acid sequences of vicilins from *Canavalia gladiata*, *C. ensiformis*, *Glycine max*, *Pisum sativum*, *Vicia faba*, *Phaseolus vulgaris*, *Ph. lunatus*, *Gossypium hirsutum*, *Theobroma cacao*, *Zea mays,* and *Picea glauca* were aligned using MACAW 2.05 (Karlin and Altschul 1990; Lawrence et al. 1993; Schuler et al. 1991). Primers were developed based on the most conserved regions of a *Ph. vulgaris* sequence (GenBank J01263). Primers 5′-TACCGTCTGTGGAGTTCAGGTCCAAACC-3′ and 5′-GAGGCCTCTAGAATATGCTTGCTGAA-3′ were used to amplify a 512-bp fragment from *Vigna luteola* genomic DNA with *Taq* DNA polymerase (Roche Applied Science). *Vigna luteola*-specific primers (viggen 5′ primer 5′-CAACCCATATTCAATACTACTACA-3′ and viggen 3′ primer 5′-GATAAAACGCAAGCATCTTATATATG-3′) amplified a 2009-bp genomic DNA fragment that includes the entire coding region, with high-fidelity polymerases: either Pwo DNA polymerase (Roche Applied Science) or Platinum PCR SuperMix High Fidelity (New England Biolabs). In some cases, the PCR products were ligated into vector pCR 2.1, and *Escherichia coli* cells were transformed with the ligation products according to the manufacturer's specifications (Invitrogen).

cDNA Synthesis and Cloning

Total RNA was extracted from mid-stage developing seeds as described by Galau et al. (1981), separated on formaldehyde agarose denaturing gels (Maniatis et al. 1982), and then used to prepare poly(A^+) RNA by adsorption to oligo d(T) cellulose (Aviv and Leder 1972). cDNA was synthesized from poly(A^+) RNA using the Smart cDNA synthesis kit (Roche Applied Science). *Eco*RI/*Not*I adaptors were ligated to the cDNA using T-4 DNA ligase (Invitrogen). The double-stranded cDNA/adaptor fragments were ligated into *Eco*RI-digested pZero and used to transform One Shot Top 10 cells following the manufacturer's directions (Invitrogen).

Isolation of Vicilin cDNA Clones

Colony filter lifts were performed as described by Grunstein and Hogness (1975). DNA was fixed to nylon membranes (Magnagraph, MSI Separations) by incubating at 85°C for 30 min or by UV irradiation with 120,000 μJ/cm^2 in a UV Stratalinker Crosslinker (Stratagene). Hybridization and washes were as recommended for dig-dUTP labeled probes (Roche Applied Science) in a Hybaid Mini Hybridization

Oven (Hybaid Instruments). The hybridized probe was detected using CSPD following the manufacturer's directions (Roche Applied Science), and chemiluminescent signals were recorded on X-ray film (Kodak X-Omat AR, Kodak Molecular Imaging Systems).

DNA Sequencing and Sequence Analysis

DNA templates were sequenced using a BigDye terminator version 1.1 cycle sequencing kit as recommended (Applied Biosystems). Nucleotide sequences were analyzed and conceptually translated using Lasergene software (DNAstar). Signal peptides and potential signal cleavage sites were identified using the SignalP program (Bendtsen et al. 2004).

Southern Hybridization

DNA was transferred and fixed to Magnagraph membranes (MSI Separations) as described (Maniatis et al. 1982). Hybridization and detection were performed as described in the Genius System User's Guide for Membrane Hybridization version 3.0 (Roche Applied Science) in a Hybaid Mini Hybridization Oven (Hybaid Instruments), washed as recommended, and the hybridized probe was detected with anti-DIG alkaline phosphatase antibody and chemiluminescent substrate (CSPD) as recommended (Roche Applied Science). Chemiluminescent signals were recorded on X-ray film (Kodak X-Omat AR, Kodak Molecular Imaging Systems).

Northern Blot Analysis

Developing seeds were classified by wet weight. Samples were ground in liquid nitrogen, and 100 mg finely ground powder was transferred to a microcentrifuge tube. Total RNA was extracted using the Spectrum Plant Total RNA Kit (Sigma-Aldrich Chemical) according to protocol A. RNA was fractionated on denaturing formaldehyde agarose gels and transferred to Magnagraph nylon membranes (MSI Separations) in 10× SSC (Maniatis et al. 1982). RNA was cross-linked to the membrane in a Stratalinker at 120 kμJ for 1 min (Stratagene). Membranes were stained 3–5 min with 0.02% methylene blue in 300 mM/l sodium acetate pH 5.5, photographed, and then destained in 0.2× SSC, 1% SDS (Sambrook and Russell 2001). Hybridization was performed in formamide buffer as recommended by the manufacturer (Roche Applied Science), in a mini-hybridization oven (Hybaid Instruments). Washes and detection of the probe with CDP-Star were as described by the manufacturer (Roche Applied Science). Chemiluminescence was detected using either a Chemi-doc System (BioRad) or Kodak X-Omat AR film (Kodak Molecular Imaging Systems).

Real-Time PCR Detection of Vicilin Genes in Population Samples

Primer and probe sequences were based on the data obtained from cloned *V. luteola* vicilin sequences (GenBank DQ060519, EU076807) with the Beacon Designer II software (Premier Biosoft). Primers were designed to span a 97-bp amplicon that included both of the sites that differentiate the two vicilin sequences (qPCR forward primer GTGAGAAGGTTGAGAAGCTGAT and qPCR reverse primer GTGAGAA GGTTGAGAAGCTGAT). Probes were dual-labeled with a 5′ fluorophore (FAM490 or HEX530) and a 3′ quencher (Black Hole 1 or 2). The sequence for the A/G probe was 56-Fam-AAGAAGCAGAGCCAATCCCACTTTGT-BHQ-1, and the sequence for the C/C probe was 5-Hex-CAGAAGCAGACCCAATCCCAC TTTGT-BHQ-2. All primers and probes were supplied by IDT (Integrated DNA Technologies). Each 25 μl qPCR reaction contained 12.5 μl iQsupermix (Bio-Rad Laboratories), 400 nM/l each primer, 300 nM/l each probe, 2 mM/l $MgCl_2$, and 25 ng genomic DNA. Plasmid DNA clones of the *V. luteola* vicilin sequences (4 pg) were used as positive controls. All the reactions were performed in duplicate on an iCycler iQ Multicolor Real-Time Detection System (Bio-Rad Laboratories) using the following amplification profile: 3 min at 95°C, followed by 60 cycles of 10 s at 5°C and 45 s at 52.1°C. Threshold cycles for both vicilin variants from each genomic sample were averaged and graphed using Sigma Plot version 10.0 (Systat Soft).

Phylogenetic Analysis of Vicilin Amino Acid Sequences

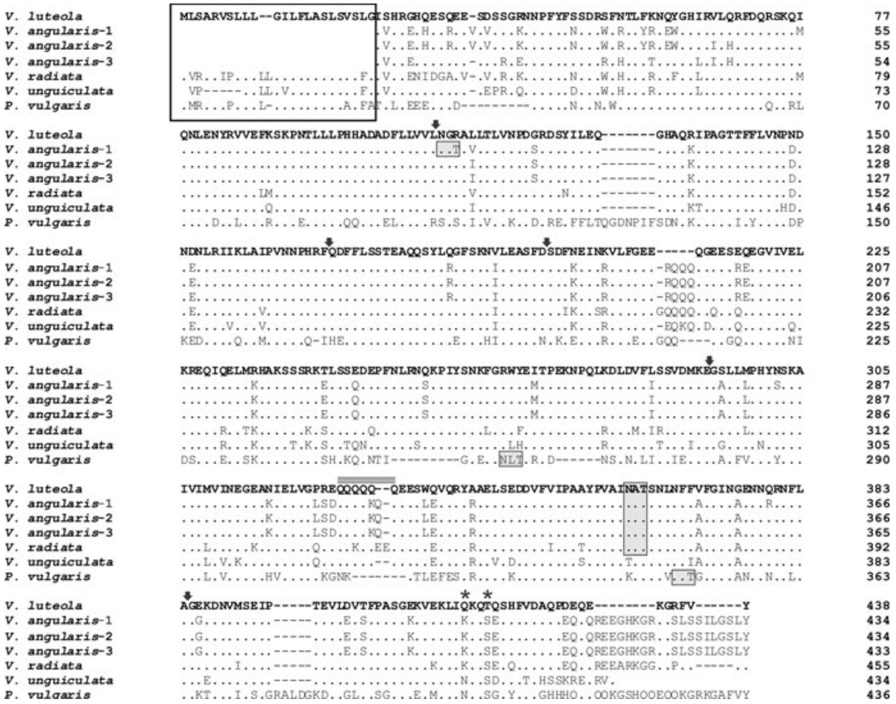
We downloaded 21 amino acid sequences of vicilins from selected legume species along with an outgroup sequence from cotton (*Gossypium hirsutum*) from the NCBI protein database (Fig. 4). Leader sequences were trimmed, and sequences were aligned with the predictive mode of Contralign 2.01 (Do et al. 2006). Next, GBlocks version 0.91 (Castresana 2000; Talavera and Castresana 2007) was used to eliminate poorly aligned and highly divergent regions, reducing the original alignment of 792 positions to a high-quality alignment of 281 positions. The online version of ProtTest (Abascal et al. 2005; http://darwin.uvigo.es/software/prottest_server.html) was then used to identify the model of sequence evolution with the highest AIC value (Akaike 1974) in a maximum-likelihood analysis. PhyML 3.0 (Guindon and Gascuel 2003) was then used to find a maximum-likelihood phylogeny with the LG+G model (Le and Gascuel 2008) identified by ProtTest, empirical estimation of amino acid residue frequencies, a gamma distribution of rates across sites with four estimated rate categories, and subtree pruning and regrafting. Branch support was assessed by a nonparametric bootstrap analysis with 1,000 replicates. The resulting phylogeny was drawn with Mega, with the sequence of *G. hirsutum* in the outgroup position and bootstrap support values scaled to 100.

Results

Characterization of the DNA and Genomic Sequences of *V. luteola* Vicilin and their Predicted Proteins

A clone encoding a vicilin seed storage protein (GenBank DQ060519) was isolated from a cDNA library generated from total RNA extracted from developing *V. luteola* seeds. Using this sequence, primers were developed that amplified the full-length genomic sequences (GenBank EU076807, EU076808). The conceptual translation product of the vicilin coding sequence from *V. luteola* has 437 amino acid residues. A BLAST search of this sequence in GenBank (release 160.0) retrieved vicilin subunit sequences with 81–87% amino acid identity from *V. radiata* and *V. angularis*, and vicilin precursors with 50–71% amino acid identity from *Arachis hypogaea*, *Canavalia ensiformis*, *C. gladiata*, *Medicago truncatula*, *Glycine max, Lupinus albus*, *Pisum sativum*, *Vicia narbonensis*, *Ph. vulgaris*, and *Lens culinaris*. The vicilin genomic sequences from *Vigna luteola* contain six exons (160–334 bp) interrupted by five small introns (88–126 bp). The introns ranked from largest to smallest are 1, 2, 5, 3, and 4. The number, relative size, and location of these introns are the same as that observed for phaseolin (GenBank J01263).

The *V. luteola* conceptual translation product was aligned with sequences from other *Vigna* species and to that of a well-characterized vicilin from *Ph. vulgaris* (Fig. 1). The predicted vicilin proteins derived from *V. luteola, V. radiata,* and two



```
V. luteola       MLSARVSLLL--GILFLASLSVSLGISHRGHQESQEE-SDSSGRNNPFYFSSDRSFNTLFKNQYGHIRVLQRFDQRSKQI   77
V. angularis-1                            .V..E.H..R..V.V...K......N...W.R..YR.EW...............M   55
V. angularis-2                            .V..E.H..R..V.V...K......N...W.R..YR.EW....I.H.........   55
V. angularis-3                            .V..E.......-...R.E..........R.H...T.....L.I.H.........   54
V. radiata       .VR..IP...LL..........F.V..ENIDGA.V-.V.R.K......N...W.H...R..F..L............M   79
V. unguiculata    VP-----..LL.V.........F.V..........-.EPR.Q......D...W.H...R.....L............   73
P. vulgaris      .MR...P...L-.........A.FAT.L.EEE...D---------.....N..N.W...................Q..RL   70

V. luteola       QNLENYRVVEFKSKPNTLLLPHHADADFLLVVLNGRALLTLVNPDGRDSYILEQ-------GHAQRIPAGTTFFLVNPND  150
V. angularis-1   ..................................T.V......S.......-------....K............D.  128
V. angularis-2   ...................................I......S.......-------....K............D.  128
V. angularis-3   ...................................I......S.......-------....K............D.  127
V. radiata       ..........LM.......................V..........N...-------....K............D.  152
V. unguiculata   ...........Q.......................I...............-------....KT..........HD.  146
P. vulgaris      ....D..L...R...E.....QQ...EL....RS.S.I.V..K..D.RE.FFLTQGDNPIFSDN.K.....I.Y....DP  150

V. luteola       NDNLRIIKLAIPVNNPHRFQDFFLSSTEAQQSYLQGFSKNVLEASFDSDFNEINKVLFGEE-----QGEESEQEGVIVEL  225
V. angularis-1   .E..........................R.....I.........K...R......-RQQQ.....RE......  207
V. angularis-2   .E..........................R.....I.........K...R......-RQQQ.....RE......  207
V. angularis-3   .E..........................R.....I.........K...R......-RQQQ.....RE......  206
V. radiata       .E........V.......................I.........IK..SR......GQQQQ..Q..Q........  232
V. unguiculata   .E....V...V.......................I.........K...R......-EQKQ.D...Q......Q.  225
P. vulgaris      KED....Q..M.....Q-IHE..............E...HI.....N.K.E...R...E..GQQ----..GQ......NI  225

V. luteola       KREQIQELMRHAKSSSRKTLSSEDEPFNLRNQKPIYSNKFGRWYEITPEKNPQLKDLDVFLSSVDMKEGSLLMPHYNSKA  305
V. angularis-1   .........K........E...Q........S.............M...............I........A..L...S...  287
V. angularis-2   .........K........E...Q........S.............M...............I........A..L......  287
V. angularis-3   .........K........E...Q........S.............M...............I........A..L......  286
V. radiata       .....R..TK......K.S.....Q.............L...F..........R...M.IR...........L......  312
V. unguiculata   .....R...K....T.K.S..TQN.....S...........LH...........R......T...I...G....N.....  305
P. vulgaris      DS...E..SK.........SH.KQ.NTI---------G.E..NLT.R.D------NS.N.LI..IE....A.FV...Y...  290

V. luteola       IVIMVINEGEANIELVGPREQQQQQ--QEESWQVQRYAAELSEDDVFVIPAAYPVAINATSNINFFVFGINGENNQRNFL  383
V. angularis-1   ...........K.....LSD....KQ-....LE....R.........................A....A...R....  366
V. angularis-2   ...........K.....LSD....KQ-....LE....R.........................A....A........  366
V. angularis-3   ...........K.....LSD....KQ-....LE....R.........................A....A........  365
V. radiata       ...L.....K.......Q....K..EE.....E....R...........I...T..............A....A........  392
V. unguiculata   ...L.V.K..........Q........--.....E....R..V.D........S.....T........IA....A.........  383
P. vulgaris      ...L.V.....HV.....KGNK-------.TLEFES.R....K..............K....V..TG....AN..N..L.  363

V. luteola       AGEKDNVMSEIP-----TEVLDVTFPASGEKVEKLIQKQTQSHFVDAQPDEQE--------KGRFV------Y  438
V. angularis-1   ..G.........-----....E.S.....K......K..SE........EQ.QREEGHKGR..SLSSILGSLY  434
V. angularis-2   ..G.........-----....E.S.....K......K..SE........EQ.QREEGHKGR..SLSSILGSLY  434
V. angularis-3   ..G.........-----....E.S.....K......K..SE........EQ.QREEGHKGR..SLSSILGSLY  433
V. radiata       .......I....-----..................K...K..SE.Q......EQ..REEARKGG..P..-----..  455
V. unguiculata   ...E........-----...................N..SD...T.HSSKRE.RV.  434
P. vulgaris      ..KT...I.S.GRALDGKD..GL..SG...E.M...N..SG.Y...GHHHO..OOKGSHOOEOOKGRKGAFVY  436
```

Fig. 1 Alignment of inferred protein sequences of five legumes. Top sequence (*bold*): *V. luteola* (GenBank EU076807). Other inferred protein sequences are *V. angularis* (GenBank AB292246, AB292247, AB292248), *V. radiata* (GenBank EF990627), *V. unguiculata* (EMBL AM905848), and a vicilin sequence from *Ph. vulgaris* (GenBank J01263). *Dash* indicates a gap in the sequence. *Dot* indicates that the amino acid is identical at the *V. luteola* site. The five intron junction positions are indicated by *arrows* above the *V. luteola* sequence. The signal peptides at the N-terminus are enclosed by a *box*, the stretch of polyglutamine residues is overscored with a *triple line*, and glycosylation sites are enclosed in *shaded boxes*. *Asterisks* at positions 415 and 418 mark the *V. luteola* vicilin sites where transitions result in amino acid substitutions

V. angularis sequences all have a putative N-linked glycosylation site at the same position in the aligned mature sequences (361–363). There is a putative N-linked glycosylation site in the predicted *Ph. vulgaris* protein close to that location (positions 341–342 of *Ph. vulgaris* in the alignment). A third *V. angularis* sequence and the *Ph. vulgaris* sequence both encode proteins with a second putative N-linked glycosylation site: for *V. angularis,* NGT at positions 89–91, and for *Ph. vulgaris,* NLT at positions 258–260 (Fig. 1).

Targeting signals direct seed storage proteins to the endoplasmic reticulum for synthesis as proproteins, which are then further modified and transported to protein storage vacuoles (Chrispeels et al. 1982). A 23-residue N-terminal signal sequence was identified in the proteins deduced from all full-length *V. luteola* vicilin sequences (Fig. 1). This sequence has 71% identity with the 24-residue signal sequence from *Ph. vulgaris*. Putative signal peptides have also been predicted for the vicilins from *V. radiata* and *V. unguiculata*. The sequences of the vicilins from *V. angularis* are truncated at the N-terminus, so signal sequences could not be identified.

A stretch of polyglutamine residues has been identified in some legume vicilins (Dure et al. 1983; Harada et al. 1989). A six-residue glutamine repeat occurs in the *V. luteola* vicilin at positions 326–331, and this repeat is modified but clearly identifiable in vicilins from the other *Vigna* species. No glutamine repeat can be identified in the vicilin of *Ph. vulgaris* in the corresponding region.

Properties of the processed vicilin peptides from these species are compared in Table 1. The mature proteins (without signal peptides) have predicted subunit molecular weights ranging from 46.7 kD (*Ph. vulgaris*) to 50.0 kD (*V. angularis*-2). The size of the predicted *V. luteola* subunit is 47.5 kD, close to the molecular weight of the vicilin subunit from *V. unguiculata*. The isoelectric point of the *V. luteola* vicilin (5.0) is the most acidic of those compared; however, all vicilin subunits compared had large net negative charges at neutral pH.

Table 1 Properties of processed vicilin subunits from *Vigna* species and a *Ph. vulgaris* vicilin

Predicted properties	Species						
	V. luteola	*V. angularis-1*	*V. angularis-2*	*V. angularis-3*	*V. radiata*	*V. unguiculata*	*Ph. vulgaris*
MW of processed subunit (kD)	47.5	49.9	50.0	49.8	49.2	47.8	46.7
Net charge at pH 7.0	−17.3	−7	−6.8	−9.8	−7.5	−13.8	−16.2
Isoelectric point	5	6	6.1	5.8	5.7	5.4	5.1
Amount of M and C (%)	1.20	1.40	1.20	1.20	0.90	0.70	0.70
Amount of N, Q, and R (%)	21.00	19.40	19.60	19.90	21.70	21.70	18.90
N-Linked glycosylation sites	1	2	1	1	1	0	2

Variation Within *V. luteola* Vicilin Genes

Only two variable sites were identified in sequences from one full-length cDNA, one partial cDNA, 10 full-length genomic vicilin sequences (cloned from a single individual), and full-length genomic sequences amplified from 10 additional individuals. Both of these variable sites occurred in exon 6 (Fig. 1). The two nucleotide differences each correspond to an amino acid replacement. A C/A transversion in the first position of codon 415 results in the replacement of a neutral glutamine residue with a positively charged lysine residue, and a C/G transversion at the second position of codon 418 results in a neutral threonine being replaced by a neutral serine residue (Fig. 1). Based on the crystallographic structures of vicilins from *V. radiata* (Itoh et al. 2006), *Ph. vulgaris* (Lawrence et al. 1990), and *C. ensiformis* (Ng et al. 1993), we predict that both of these variable sites lie within the third alpha-helical region of the C-terminus of the protein.

We are able to detect both of these variant *V. luteola* vicilin gene transcripts in cDNA from developing seeds of *V. luteola.* Therefore, both of these genes are expressed and are not pseudogenes (data not shown). The two different vicilin gene sequences (GenBank EU076807, DQ060520) are not allelic variants at a single locus. Sequence-specific dual-labeled fluorogenic probes (Taqman probes) detected both of these sequences in every one of 72 individual plants collected from southern Louisiana (Fig. 2). If these were allelic variants in a sexually reproducing population, we would expect at least half of the individuals to be homozygous

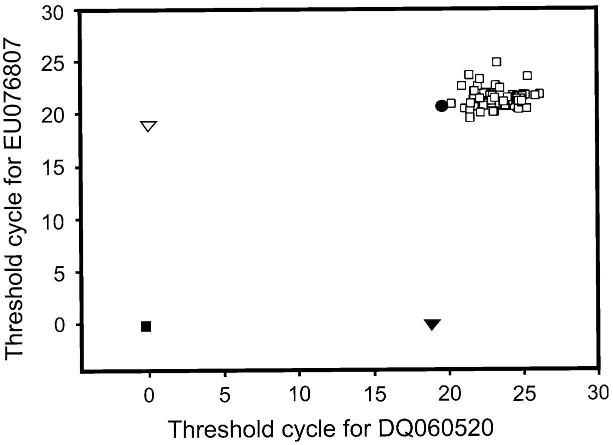


Fig. 2 Distribution of GenBank EU076807 and DQ06520 sequences in populations of *V. luteola,* detected by real-time PCR. *Black circle* 1:1 Mixture of GenBank EU076807 and DQ06520 (full-length genomic DNAs cloned into PCR 2.1). *Inverted open triangle* Pure sample of GenBank EU076807 clone. *Inverted black triangle* Pure sample of GenBank DQ06520 clone. *Open squares* DNA extracted from *V. luteola* individuals (Chlan lab accession numbers CC4, CC13, CC26, CC33, CC67, CC00641–CC00649, CC00651–CC00654, CC00670–CC00679, CC00681–CC00689, CC00691, CC006410–CC006421, CC006710–CC006714, CC006718, CC006720, CC006722, CC006724, CC006726–CC006728, CC006810, CC006811, CC006813-CC006816, CC006818, CC006819, CC006821–CC006823) collected from different populations. *Black square* Negative control: genomic DNA from *Rhynchosia minima* (Chlan lab accession number CC12)

for either one of the variants. The presence of two sequences in all individuals indicates the presence of two vicilin loci. Vicilin gene duplications have been previously reported (Chlan et al. 1987; Slightom et al. 1983; Talbot et al. 1984).

High levels of interspecific divergence in vicilin sequences could be explained by reduced purifying selection, which would also allow the accumulation of high levels of intraspecific polymorphism; however, we found no polymorphism in either of the vicilin genes of *V. luteola*. It is possible that polymorphism has been eliminated by a recent genetic bottleneck. Analysis of variation at other loci will be used to examine this possibility in the future. It is also possible that interspecific divergence in vicilin amino acid sequences is a response to directional selection rather than reduced purifying selection. We plan to investigate this possibility once we have sequence information from other wild legume vicilin genes.

Vigna luteola Vicilin Gene Copy Number

Southern hybridization analysis of genomic DNA (Fig. 3) also provides evidence for duplication of the vicilin gene in *V. luteola*. Based on the intensity and number of bands, we estimate that there are four copies of vicilin sequences within the diploid genome (Bennett and Leitch 1997) of *V. luteola*, or two copies per haploid genome. Because these sequences differ at only two nucleotide positions, it appears that the gene duplication event was recent.

The duplication of the vicilin locus in *V. luteola* is not unusual. Gene duplications were discovered in early studies of seed storage genes, which led to the impression that these genes evolved by unusual mechanisms (Doyle et al. 1986). More recent analyses of complete genomes have revealed that many plant genes have undergone

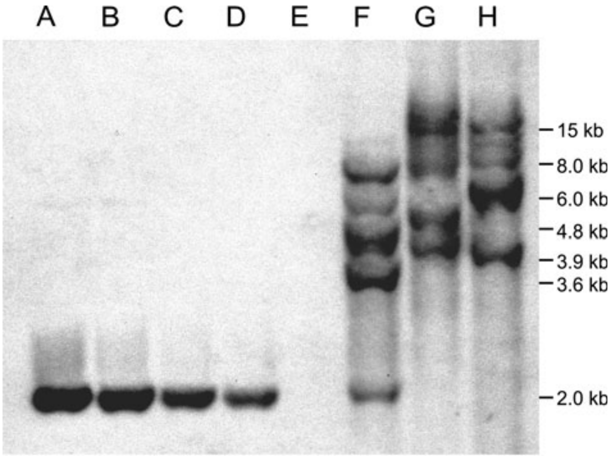


Fig. 3 Southern blot analysis of *V. luteola* genomic DNA. The *first five lanes* contain 8 (*lane A*), 4 (*lane B*), 2 (*lane C*), 1 (*lane D*), and 0 (*lane E*) copy number controls constructed using a cloned genomic DNA sequence (GenBank EU76807). In the remaining lanes, 5.0 µg of *V. luteola* genomic DNA was digested with enzymes that do not have restriction sites in the probe sequence: *Hin*dIII (*lane F*), *Bgl*II (*lane G*), *Eco*RV (*lane H*)

duplication and form multigene families (Sappl et al. 2004). It is interesting that the nucleotide substitutions that differentiate the two vicilin loci in *V. luteola* both result in amino acid replacements. This could indicate a lack of purifying selection at these sites, although both loci are transcribed in seed tissue and are presumably subjected to functional constraints.

Amino Acid Composition of *V. luteola* Vicilins Compared with Other Legume Vicilins

We compared the amino acid composition of the vicilins predicted from gene sequences of *V. luteola* with those of other legume vicilins. The molar percentage of sulfur-containing amino acids in *V. luteola* vicilin differs from the averages found in legume 7S proteins from other genera. Typically, legume vicilins contain relatively low proportions of sulfur-containing amino acid residues. We determined the amino acid averages for 32 legume vicilin sequences and compared those values for *V. luteola* (Table 2). The average amount of cysteine and methionine together in the legume vicilins we compared is less than 0.7% of the total amino acid content, versus 1.2% in *V. luteola*. Surprisingly, in *V. luteola*, all of the sulfur is contributed by methionine; there are no cysteine residues. Predicted proteins of vicilins of *V. angularis* (Fukuda et al. 2007), *V. radiata* (Bernardo et al. 2004), and *V. unguiculata* (GenBank AM905848) also contain no cysteine, and neither does the *Ph. vulgaris* vicilin. There is a difference between the total percentage of methionine residues in these vicilins. In one of the isoforms of *V. angularis,* 1.4% of the amino acid residues are methionine, and the percentage of met residues in the other two isoforms and *V. luteola* is 1.2%. *Vigna unguiculata* and *Ph. vulgaris* have the lowest percentage (0.7%). The vicilins in the genus *Vigna* that have been identified so far all lack cysteine residues in the processed proteins. Many legume vicilins, however, do contain cysteine residues, either in the leader peptide or in the mature protein. For example, *Glycine max* has a cysteine residue in the leader peptide (Harada et al. 1989), and the mature vicilin subunit in *Canavalia ensiformis* contains two cysteine residues (Ng et al. 1993).

Another common characteristic of vicilin amino acid composition is the high percentage of nitrogen-enriched amino acids (glutamine, asparagine, and arginine). In our comparison of 32 vicilins from several legumes (Table 2), the average molar percentage of asparagine residue was 7.4%, glutamine was 7.2%, and arginine was 6.4%. In vicilins of *V. luteola*, the molar percentages of glutamine and asparagine are similar to the averages for the vicilins of species in our comparison (8.0% in both). The 5.1% molar percentage of arginine in vicilins of *V. luteola* was less than the average for all the vicilins of legumes that we compared (6.4%). The amino acid composition of vicilin from *V. luteola* has a few other distinctive characteristics. The vicilins from *V. luteola* contain fewer lysine and arginine residues than aspartic acid and glutamic acid residues, which accounts for their acidic pI. Compared with the averages we observed for vicilins in general, the *V. luteola* vicilins have higher molar percentages of the nonpolar amino acids phenylalanine, valine, and tryptophan.

Table 2 Amino acids in 32 vicilin sequences from legumes

Amino acid	Average % of 10 species[a]	*V. luteola* %[b]
Glutamine	7.2 ± 1.2	8.0
Asparagine	7.4 ± 0.9	8.0
Arginine	6.4 ± 1.7	5.1
Lysine	6.2 ± 0.9	5.1
Aspartate	4.9 ± 0.5	5.3
Glutamate	11 ± 2.4	9.4
Phenylalanine	5.1 ± 0.9	**6.5**
Valine	5.6 ± 0.9	**6.8**
Tryptophan	0.2 ± 0.2	**0.5**
Methionine	0.4 ± 0.4	**1.2**
Cysteine	0.3 ± 0.3	0
Leucine	9.3 ± 1.4	8.4
Isoleucine	5.5 ± 0.8	5.3
Alanine	4.8 ± 0.7	4.8
Serine	8.3 ± 0.8	8.3
Glycine	5.2 ± 0.7	4.8
Proline	4.7 ± 1.0	4.6
Tyrosine	2.8 ± 0.7	2.7
Threonine	2.7 ± 0.6	2.9
Histidine	2.1 ± 0.7	2.4

[a] Molar percentage of amino acid averaged among vicilins from 10 species of legumes: *Canavalia ensiformis*, *C. gladiata*, *Glycine max*, *Lens culinaris*, *Lupinus albus*, *Pisum sativum*, *Phaseolus lunatus*, *Ph. vulgaris*, *Vicia faba,* and *Vicia narbonensis*

[b] Molar percentage for the predicted amino acid sequence of the *V. luteola* vicilin. Bold values are outside the range for the 10 averaged species

Phylogenetic Analysis of Legume Vicilin Domains

Phylogenetic analysis of the amino acid sequences of vicilin sequences of *V. luteola* and other representative legume species was performed and a phylogenetic tree constructed (Fig. 4). The vicilin sequences from all the *Vigna* species analyzed formed a well-supported sister group to vicilin sequences of *Ph. vulgaris* and *Ph. lunatus*. This finding is consistent with current classifications of legumes (Doyle et al. 2000). The *V. luteola* vicilin is placed on a separate branch from the other *Vigna* vicilins, but the bootstrap value for this node is low (45). The results are also consistent with vicilin gene duplication and subsequent divergence within the Vicieae. For example, *Vicia narbonensis*-1 clusters with *Pisum sativum*-1, and *Vicia narbonensis*-2 clusters with *Pisum sativum*-2. These groupings are supported by bootstrap values of 100 and 67, respectively.

Expression of *V. luteola* Vicilins

Transcription of vicilin mRNA in *V. luteola* is limited to developing seeds (Fig. 5), which is typical for vicilin expression in plants. Expression is also temporally regulated (Fig. 5). Vicilin mRNAs were not detected in very young developing seeds, reached high levels in the middle stages of seed development, decreased during the later stages, and were undetectable in mature seeds. This time course of expression differs from that reported for *Medicago truncatula*, in which vicilin

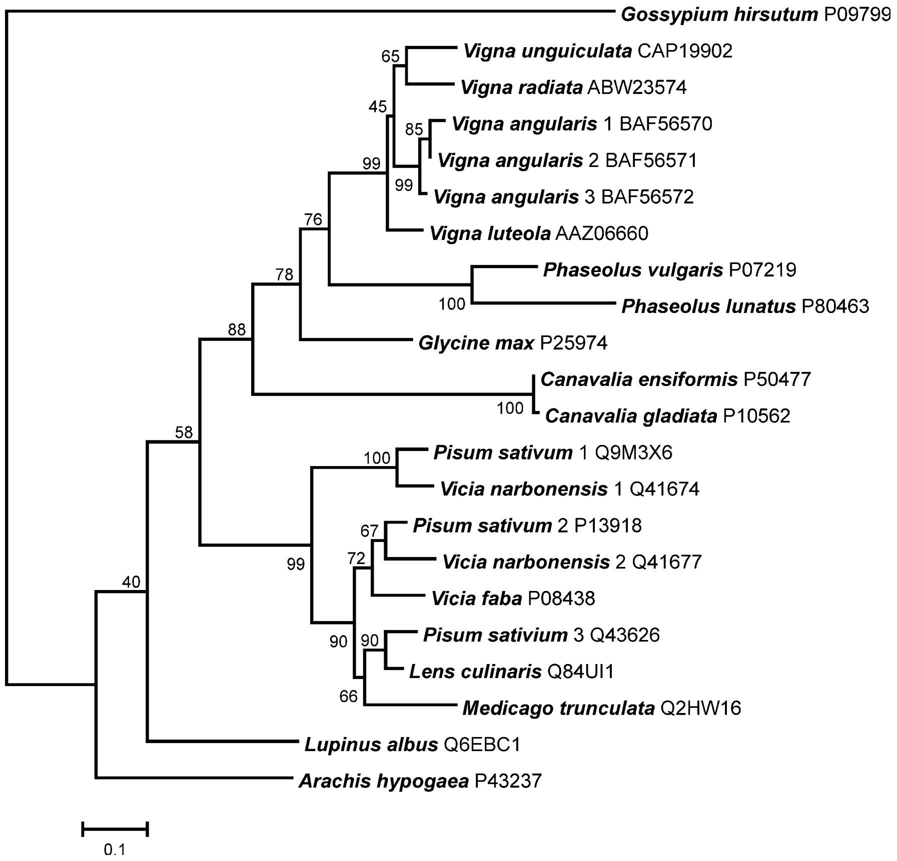

Gossypium hirsutum P09799
Vigna unguiculata CAP19902
Vigna radiata ABW23574
Vigna angularis 1 BAF56570
Vigna angularis 2 BAF56571
Vigna angularis 3 BAF56572
Vigna luteola AAZ06660
Phaseolus vulgaris P07219
Phaseolus lunatus P80463
Glycine max P25974
Canavalia ensiformis P50477
Canavalia gladiata P10562
Pisum sativum 1 Q9M3X6
Vicia narbonensis 1 Q41674
Pisum sativum 2 P13918
Vicia narbonensis 2 Q41677
Vicia faba P08438
Pisum sativium 3 Q43626
Lens culinaris Q84UI1
Medicago trunculata Q2HW16
Lupinus albus Q6EBC1
Arachis hypogaea P43237
65
45
85
99
99
76
78
100
88
100
58
100
67
99
72
90
90
66
40
0.1

Fig. 4 Phylogenetic relationships of the vicilin sequences of *V. luteola* and other representative legume species. GenBank accession number of each sequence follows the species name in the maximum-likelihood tree, with bootstrap support values at each branch (1,000 replicates). Cotton (*Gossypium hirsutum*) was the outgroup

proteins accumulated dramatically during the very early stages of seed development and became less abundant during the later stages of development (Gallardo et al. 2003). The pattern of vicilin mRNA accumulation in *V. luteola* is more similar to that seen in *C. gladiata* (Yamauchi et al. 1988).

Discussion

The vicilin genes in *V. luteola* are expressed only in developing seeds. Vicilin proteins have high levels of sulfur-containing and nitrogen-rich amino acids relative to vicilins of other taxa. There are no cysteine residues in the predicted mature vicilins of *V. luteola*; the only sulfur-containing amino acid is methionine.

Two copies of the vicilin gene have been identified in the wild legume *V. luteola*. They are similar to the vicilin genes of other species in structure and in

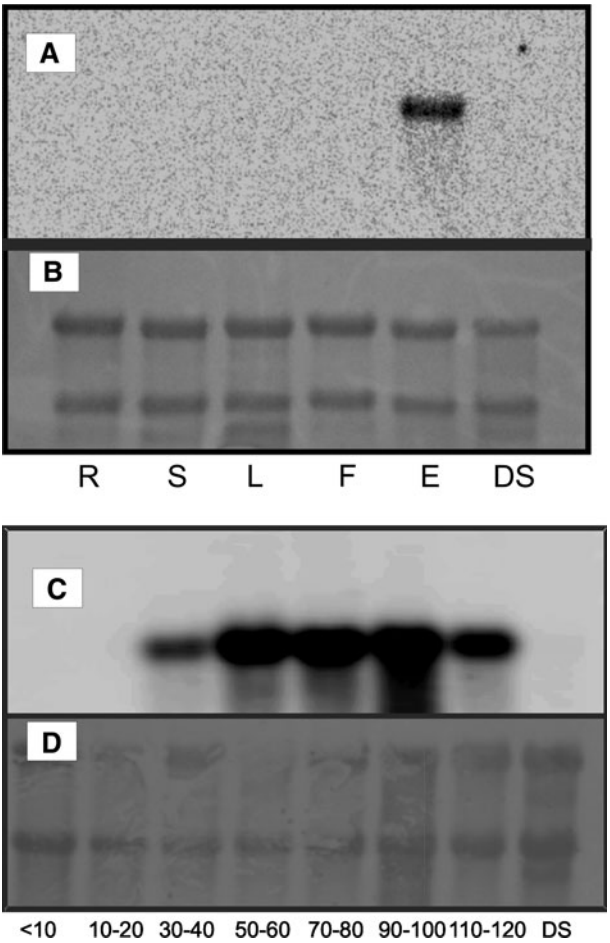


Fig. 5 Northern blot analysis of vicilin mRNA expression patterns. **A**, **B** Tissue specificity of vicilin expression in *V. luteola,* limited to developing seeds: Total RNA (5 μg) from *V. luteola* root (R), stem (S), leaf (L), flower (F), 60 mg embryo (E), and dry seed (DS) tissues was fractionated on denaturing agarose gels, transferred to a nylon filter, and hybridized with a *V. luteola* vicilin probe. Before hybridization, the filter was stained with methylene blue as a loading and transfer control. **C**, **D** Developmental expression of vicilin mRNA in *V. luteola:* Total RNA (5 μg) from developing *V. luteola* seeds at eight stages of development (*left to right:* <10 mg, 10–20 mg, 30–40 mg, 50–60 mg, 70–80 mg, 90–100 mg, 110–120 mg, and dry seeds) were fractionated and stained as described for **B**

characteristic features, which include a signal sequence and a stretch of repeated glutamine residues. The two copies of the vicilin gene that we have identified in *V. luteola* differ at only two nucleotide sites, suggesting a recent gene duplication. Analysis of vicilin gene sequences from 11 individuals isolated from six wild populations of *V. luteola* did not reveal any genetic polymorphisms. Additional studies to characterize vicilin gene sequence variation within wild populations of legumes and sequence divergence between legume species are ongoing. Results from these studies will be used to test hypotheses about the roles of different forms of selection in the evolution of vicilin genes and proteins.

The organization and expression patterns observed for the vicilins from *V. luteola* are similar to those observed for other legume vicilins. In a comparison with a vicilin from *Ph. vulgaris*, the number of introns is identical and the locations of intron splice sites are similar. The tissue specificity and developmental regulation of their expression is similar to that seen for other vicilins. The *V. luteola* vicilin protein subunits do have some unusual properties. First, the percentage of sulfur-containing amino acids in the predicted mature protein is relatively high (1.2%) among comparable species (0.7–1.4% for other *Vigna* vicilins and 0.7% for *Ph. vulgaris;* Table 1). It is also higher than the percentage in mature vicilin subunits of *Vicia faba* (0%), *Pisum sativum* (0.3%), and *Glycine max* (0%). Second, the predicted mature protein does not contain any cysteine residues. Vicilin-like sequences from *Vigna angularis*, *V. radiata,* and *V. unguiculata* also do not encode predicted proteins with cysteine residues. Other legume vicilins, such as those from *Canavalia ensiformis* and *C. gladiata,* do contain cys residues in the mature protein, and some other legume vicilin proteins contain a single cys residue that is located in the leader peptide (*Vicia narbonensis*, *Vicia faba*, *Glycine max*, and *Pisum sativum*). The lack of cys residues precludes formation of disulfide bonds, which could potentially affect vicilin protein structure or function. The 3D structure has been determined for *Ph. vulgaris* vicilin chain A (Ko et al. 1993, 2000). Although, this protein lacks cys residues, it forms alpha-helical and beta barrel structures similar to those identified in the cys-containing *C. ensiformis* vicilin protein (Lawrence et al. 1990). It is therefore unlikely that cys residues play a critical role in the formation of canonical vicilin subunit structures. Cys residues in some vicilins are associated with antimicrobial activity. A repeating cys-containing motif (cys-x$_3$-cys–×10–15-cys–x$_3$-cys) has been identified in several vicilins (Chlan et al. 1986; Marcus et al. 1999). Peptides that contain these repeated cys motifs exhibit antimicrobial activity in vitro (Marcus et al. 1999). Disulfide bonds may form between these repeats and stabilize the structure (Borroto and Dure 1987). In the absence of cys residues, as in the case of *V. luteola* vicilin, degradation products would not be stabilized by disulfide bonds, which could affect their half-life and potential activity.

The most surprising finding is this study is that the two vicilin sequences that we characterized in *V. luteola* differ by only two nucleotides. No polymorphic sites were detected in either gene in our analysis of multiple individuals from different populations. In contrast, the vicilin sequences of *V. angularis* differ by as much as 2.3%. This lack of divergence observed for the *V. luteola* sequences could be explained by either a recent gene duplication or a recent bottleneck.

We are interested in determining what forces or constraints influence vicilin structural conservation and function. Clearly, vicilin structure is extremely important to its function; vicilins are not simple junk proteins that can include random amino acid sequences and still function. As a first step in this process, we have characterized the vicilin gene structure, expression, and variation within *V. luteola,* a wild legume that is widely distributed in the southern United States.

We have shown that although there are similarities between *V. luteola* vicilins and other legume vicilins, the vicilins of *V. luteola* are unique in some aspects. Comparative studies of the structure and properties of the vicilins of additional wild legume populations will enable us to gain insight into the evolution of these important genes in a genetically diverse population.

Acknowledgments We thank Ms. Angela Geraci and Mr. Vijay Chigarapati for critical reading of this manuscript. We thank Dr. Joel Stake and Dr. Brian Mahon for assistance in collection of field samples. This study was supported by a State of Louisiana Board of Regents Support Fund Research and Development Grant LEQSF (2002-04)-RD-A-32.

References

Abascal F, Zardoya R, Posada D (2005) ProtTest: selection of best-fit models of protein evolution. Bioinformatics 21:2104–2105

Akaike H (1974) A new look at statistical model identification. IEEE Trans Autom Control AC 19:716–723

Aviv H, Leder P (1972) Purification of biologically active globin messenger RNA by chromatography on oligothymidylic acid-cellulose. Proc Natl Acad Sci USA 134:743

Bendtsen JD, Nielsen H, Heijne GV, Brunak S (2004) Improved prediction of signal peptides: SignalP 3.0. J Mol Biol 340:783–795

Bennett MD, Leitch IJ (1997) Nuclear DNA amounts in angiosperms: 583 new estimates. Ann Bot (Lond) 80:169–196

Bernardo AE, Gracia RN, Adachi M, Angeles JG, Kaga A, Ishimoto M, Utsumi S, Tecson-Mendoza EM (2004) 8S globulin of mungbean [*Vigna radiata* (L.) Wilczek]: cloning and characterization of its cDNA isoforms, expression in *Escherichia coli*, purification and crystallization of the major recombinant 8S isoform. J Agric Food Chem 52:2552–2560

Borroto K, Dure LS III (1987) The globulin seed storage proteins of flowering plants are derived from two ancestral genes. Plant Mol Biol 8:113–131

Bown D, Ellis THN, Gatehouse JA (1988) The sequence of a gene encoding convicilin from pea (*Pisum-Sativum*-L.) shows that convicilin differs from vicilin by an insertion near the N-terminus. Biochem J 251:717–726

Castresana J (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol 17:540–552

Chlan CA, Pyle JB, Legocki AB, Dure LS III (1986) Developmental biochemistry of cottonseed embryogenesis and germination, 18: cDNA and amino acid sequences of members of the storage protein families. Plant Mol Biol 7:475–489

Chlan CA, Borroto K, Kamalay JA, Dure LS III (1987) Developmental biochemistry of cottonseed embryogenesis and germination, 19: sequences and genomic organization of the α globulin (vicilin) genes of cottonseed. Plant Mol Biol 9:533–546

Chrispeels MJ, Higgins TJV, Spencer D (1982) Assembly of storage protein oligomers in the endoplasmic reticulum, and processing of the polypeptides in the protein bodies of developing pea cotyledons. J Cell Biol 93:306–313

Do CB, Gross SS, Batzoglou S (2006) CONTRAlign: discriminative training for protein sequence alignment. In: Apostolico A, Guerra C, Istrail S, Pevzner P, Waterman M (eds) Research in computational molecular biology. Springer, Heidelberg, pp. 160–174

Doyle JJ, Schuler MA, Godette WD, Zenger V, Beachy RN, Slightom JL (1986) The glycosylated seed storage proteins of *Glycine max* and *Phaseolus vulgaris:* structural homologies of genes and proteins. J Biol Chem 261:9228–9238

Doyle JJ, Chappill JA, Bailey DC, Kajita T (2000) Towards a comprehensive phylogeny of legumes: evidence from rbcL sequences and non-molecular data. In: Herendeen PS, Bruneau A (eds) Advances in legume systematics. Royal Botanic Gardens, Kew

Dure LS III, Pyle JB, Chlan CA, Baker JC, Galau GA (1983) Developmental biochemistry of cottonseed embryogenesis and germination, 17: developmental expression of genes for the principal storage proteins. Plant Mol Biol 2:199–206

Fukuda T, Prak K, Fujioka M, Maruyama N, Utsumi S (2007) Physicochemical properties of native adzuki bean (*Vigna angularis*) 7S globulin and the molecular cloning of its cDNA isoforms. J Agric Food Chem 55:3667–3674

Galau GA, Legocki AB, Greenway SC, Dure LS III (1981) Cotton messenger RNA sequences exist in both polyadenylated and nonpolyadenylated forms. J Biol Chem 256:2551–2560

Gallardo K, Le Signor C, Vandekerckhove J, Thompson RD, Burstin J (2003) Proteomics of *Medicago truncatula* seed development establishes the time frame of diverse metabolic processes related to reserve accumulation. Plant Physiol 133:664–682

Grunstein M, Hogness DS (1975) Colony hybridization: a method for the isolation of cloned DNAs that contain a specific gene. Proc Natl Acad Sci USA 72:3961–3965

Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol 52:696–704

Harada JJ, Barker SJ, Goldberg RB (1989) Soybean β-conglycinin genes are clustered in several DNA regions and are regulated by transcriptional and posttranscriptional processes. Plant Cell 1:415–425

Itoh T, Garcia RN, Adachi M, Maruyama Y, Tecson-Mendoza EM, Mikami B, Utsumi S (2006) Structure of 8S alpha globulin, the major seed storage protein of mung bean. Acta Cryst Sect D62:824–832

Karlin S, Altschul SF (1990) Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. Proc Natl Acad Sci USA 87:2264–2268

Ko TP, Ng JD, McPherson A (1993) The three-dimensional structure of canavalin from jack bean (*Canavalia ensiformis*). Plant Physiol 101:729–744

Ko TP, Day J, McPherson A (2000) The refined structure of canavalin from jack bean in two crystal forms at 2.1 and 2.0 Å resolution. Acta Cryst Sect D 56:411–420

Lawrence MC, Suzuki E, Varghese JN, Davis PC, Vandonkelaar A, Tulloch PA, Colman PM (1990) The 3-dimensional structure of the seed storage protein phaseolin at 3 Å resolution. EMBO J 9:9–15

Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, Wootton JC (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. Science 262:208–214

Lawrence MC, Lzard T, Beuchat M, Blagrove RJ, Colman PM (1994) Structure of phaseolin at 2.2 Å resolution: implications for a common vicilin/legumin structure and the genetic engineering of seed storage proteins. J Mol Biol 238:748–776

Le SQ, Gascuel O (2008) An improved general amino acid replacement matrix. Mol Biol Evol 25:1307–1320

Li H, Luo J, Hemphill JK, Juang J-T, Gould J (2001) A rapid and high-yielding DNA miniprep for cotton (*Gossypium* spp.). Plant Mol Biol Rep 19:1–5

Maniatis T, Fritsch E, Sambrook J (1982) Molecular cloning: a laboratory manual. Cold Spring Harbor Laboratory, Cold Spring Harbor, NY

Marcus JP, Green JL, Gouter KC, Manners JM (1999) A family of antimicrobial peptides is produces by processing of a 7S globulin protein in *Macadamia integrifolia* kernals. Plant J 19:699–710

Ng JD, Ko TP, McPherson A (1993) Cloning, expression, and crystallization of jack bean (*Canavalia ensiformis*) canavalin. Plant Physiol 101:713–728

Osborne TB (1924) The vegetable proteins, 2nd edn. Longmans Green and Co., London

Paterson AH, Brubaker CL, Wendel JF (1993) A rapid method for extraction of cotton (*Gossypium* spp.) genomic DNA suitable for RFLP or PCR analysis. Plant Mol Biol Rep 11:122–127

Pernollet J-C, Mossé J (1983) Structure and location of legume and cereal seed storage proteins. Academic Press, New York, NY

Sambrook J, Russell DW (2001) Molecular cloning: a laboratory manual, 3rd edn. Cold Spring Harbor Press, New York, NY

Sappl PG, Heazlewood JL, Millar AH (2004) Untangling multigene families in plants by integrating proteomics into functional genomics. Phytochemistry 65:1517–1530

Schuler GD, Altschul SF, Lipman DJ (1991) A workbench for multiple alignment construction and analysis. Struct Funct Genet 9:180–190

Shewry PR, Halford NG (2002) Cereal seed storage proteins: structures, properties and role in grain utilization. J Exp Bot 53:947–958

Slightom JL, Sun SM, Hall TC (1983) Complete nucleotide sequence of a French bean storage protein gene: phaseolin. Proc Natl Acad Sci USA 80:1897–1901

Talavera G, Castresana J (2007) Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst Biol 56:564–577

Talbot DR, Adang MJ, Slightom JL, Hall TC (1984) Size and organization of a multigene family encoding phaseolin, the major seed storage protein of *Phaseolus vulgaris* L. Mol Gen Genet 198:42–49

Viques OM, Konan KN, Dodo HW (2003) Structure and organization of the genomic clone of a major peanut allergen gene, Ara h 1. Mol Immunol 40:565–571

Yamauchi D, Nakamura K, Asahi T, Minamikawa T (1988) cDNAs for canavalin and concanavalin A from *Canavalia gladiata* seeds: nucleotide sequence of cDNA for canavalin and RNA blot analysis of canavalin and concanavalin A mRNAs in developing seeds. Eur J Biochem 170:515–520