Biochem. J. (1988) 250, 15–24 (Printed in Great Britain)



# Two genes encoding 'minor' legumin polypeptides in pea (*Pisum sativum* L.)

## Characterization and complete sequence of the *LegJ* gene

John A. GATEHOUSE,*‡ David BOWN,* John GILROY,* Mark LEVASSEUR,* Judy CASTLETON† and T. H. Noel ELLIS†

*Department of Botany, Durham University, South Road, Durham DH1 3LE, U.K., and †John Innes Institute, Colney Lane, Norwich NR4 7UH, U.K.

A genomic clone from pea (*Pisum sativum* L.) contains all of one gene encoding a 'minor' (B-type) legumin polypeptide, and most of a second very similar gene. The two genes, designated *LegJ* and *LegK*, are arranged in tandem, separated by approx. 6 kb. A complete sequence of gene *LegJ* and its flanking sequences is given, with as much of the sequence of gene *LegK* as is present on the genomic clone. Hybridization of 3′ flanking sequence probes to seed mRNA, and sequence comparisons with cDNA species, suggested that gene *LegJ*, and probably gene *LegK*, was expressed. The partial amino acid sequences of 'minor' legumin α- and β-polypeptides were used to confirm the identity of these genes. The transciption start in gene *LegJ* was mapped. The 5′ flanking sequence of gene *LegJ* contains a sequence conserved in legumin genes from pea and other species, which is likely to have functional significance in control of gene expression. Sequence comparisons with legumin genes and cDNA species from *Vicia faba* and soya bean show that separation of legumin genes into A- and B-type subfamilies occurred before separation of the Viciae and Glycinae tribes.

## INTRODUCTION

The seeds of pea (*Pisum sativum* L.) contain several types of heterogeneous proteins that function as storage reserves to be utilized on seed germination. Legumin is one of these storage protein types; it contains molecules made up of six disulphide-bonded subunit pairs, each subunit pair (polypeptides of $M_r$ approx. 40000 and 20000) being synthesized as a single precursor polypeptide ($M_r$ approx. 60000) [1,2]. Homologous legumin-type proteins are found in the seeds of many other plant species, both legumes and non-legumes [3,4]. In pea, the subunit pair types found in legumin may be conveniently divided into 'major' components and 'minor' components [5] on the basis of their relative abundances in the protein. The 'major' legumin subunit pairs are encoded by a small gene family containing five members [6,7]; several of these genes have been fully characterized and sequenced ([8,9]; A. H. Shirsat, S. Mahmoud, R. R. D. Croy & J. A. Gatehouse, unpublished work).

Studies of cDNA libraries based on mRNA isolated from developing pea seeds have shown that, besides a class of cDNA species that encode 'major' legumin subunit pairs and correspond to the *LegA* family of genes, two further classes of legumin cDNA species are present. These additional cDNA species do not cross-hybridize at the nucleic acid level to *LegA*-type cDNA species, but encode polypeptides that are immunoprecipitable by anti-legumin antibodies in hybrid-release translation experiments [10]. The encoded polypeptides are thus thought to be the precursors of 'minor' legumin subunit pairs; support for this hypothesis has been given by partial nucleotide sequences of these cDNA species [11]. Hybridization of these 'minor' legumin cDNAs to pea genomic DNA has shown that a further four or five legumin genes belong to this second family, which can be divided into two sub-families (containing three and one or two genes) on the basis of hybridization homology of the different 'minor' legumin cDNA species [7].

In the present paper one of the 'minor' legumin cDNA species was used as a probe towards a pea genomic library, in order to isolate two genes from its corresponding sub-family.

## MATERIALS AND METHODS

### Materials

Pea seeds cultivar Feltham First were obtained from Suttons Seeds, Torquay, Devon, U.K. The isolation of the genomic clone λ JC5 from a *Pisum* genomic library (prepared by ligation of a size-fractionated *Eco*RI partial digest of DNA from pea cultivar Dark Skinned Perfection to the bacteriophage vector λ EMBL4, packaging, and transfection of recombinant bacteriophage) by screening the library with a cDNA species, pCD40, that had been shown to hybrid-select 'minor' legumin precursor polypeptides of $M_r$ 63000–65000 from a translation *in vitro* of pea seed mRNA [10] has been described elsewhere [12]. Restriction enzymes were from Northumbrian Biologicals, Cramlington, Northumberland, U.K.; S1 nuclease and other enzymes were from BCL, Lewes, E. Sussex, U.K. Reagents and enzymes for M13 DNA sequencing were supplied by Gibco/BRL

‡ To whom correspondence should be addressed.

These sequence data have been submitted to the EMBL/GenBank Data Libraries under the accession number Y00715.

(Gibco, Paisley, Strathclyde, Scotland, U.K.); non-standard sequencing primers were synthesized on an Applied Biosystems model 381A DNA synthesizer. Radiochemicals were from Amersham International, Amersham, Bucks., U.K., as were the reagents for nick translation. The random hexanucleotide-primer labelling reagents (Polymeraid) were supplied by P & S Biochemicals, Liverpool, U.K. Nitrocellulose filters were Schleicher and Schuell type BA85 from Andermann and Co., East Molesey, Surrey, U.K. Other reagents were of analytical quality or best available grade.

### Methods

**DNA subcloning and sequencing.** Subcloning of appropriate restriction fragments from λ JC5 into the plasmid vector pUC9 and into M13 vectors followed standard protocols [13]. DNA probes were labelled with [α-$^{32}$P]-dCTP (400 Ci/mmol; 100 μCi used per 0.2–0.5 μg of DNA) by nick translation [14] except where otherwise stated. All DNA sequencing was carried out by the dideoxynucleotide chain-termination method [15] with single-stranded DNA prepared from cloned fragments in M13 mp8, M13 mp9, M13 mp18 or M13 mp19 [16], with [α-[$^{35}$S]thio]dATP as radioactive label [17]. DNA sequences were read and analysed manually; standard computer programs were used for translation and dot-matrix sequence comparison [18]. DNA sequences were also compared by the dimensional plot method [19], by using a computer program devised by the authors.

**S1-nuclease mapping.** S1-nuclease mapping of the 5′ end of the mRNA species hydribizing to the *LegJ* gene was carried out according to the method of Favaloro *et al.* [20] with *Sph*I–*Msp*I and *Sph*I–*Xho*I DNA probe fragments from gene *LegJ*, 5′-end-labelled with [γ-$^{32}$P]ATP (6000 Ci/mmol; 50 μCi used per 0.2–0.5 μg of DNA) [21]. Labelled fragments (at least $10^6$ c.p.m. per assay) were hybridized to polyadenylated RNA purified from developing pea cotyledons [22] (5 μg per assay). Protected fragments after S1-nuclease digestion were sized by electrophoresis on DNA sequencing gels.

**Hybridization to mRNA.** Restriction fragments from genes *LegJ* and *LegK* were isolated by excision of bands from low-melting-temperature agarose gels after separation of the fragments by electrophoresis [23], and were labelled with [α-$^{32}$P]dCTP (400 Ci/mmol; 50 μCi used per 25 ng of DNA) by using the random hexanucleotide-primer method [24]. Total RNA preparations were purified from developing pea cotyledons (cultivar. Feltham First) by extraction in guanidinium thiocyanate as previously described [25]. RNA samples were denatured with glyoxal and analysed by electrophoresis on 1.2% agarose gels [26]. The gels were blotted on to nitrocellulose, and the blots were hybridized with labelled probes in a hybridization solution containing 5 × SSC (1 × SSC is 0.15 M-NaCl/0.015 M-sodium citrate buffer, pH 7.2), 2 × Denhardt's solution (1 × Denhardt's solution is 0.02% Ficoll/0.02% bovine serum albumin/0.02% polyvinylpyrrolidone), 200 μg of denatured herring sperm DNA/ml and 50% formamide at 42 °C, as described by Thomas [27]. The blots were washed to a final stringency of 0.1 × SSC/0.1% SDS at 50 °C, before autoradiography at −80 °C with a preflashed X-ray film (Fuji RX) and an intensifying screen (Cronex Lightning Plus; DuPont).

**Protein purification and sequencing.** A 'minor' legumin polypeptide pair, corresponding to that designated L2 by Matta *et al.* [28], was isolated from pea legumin (purified as described in ref. [29]) by ion-exchange chromatography on a column of DEAE-Sepharose in 50 mM-Tris/HCl buffer, pH 7.5, containing 6 M-urea, essentially as described by Casey *et al.* [30]. The column was eluted by a salt gradient, the L2 polypeptide pair being eluted at [NaCl] approx. 0.3 M. Pooled fractions were reduced and carboxymethylated with iodoacetic acid, and the α- and β-polypeptides of L2 were then separated as indicated above. Polypeptides were subjected to tryptic digestion, and the resultant peptides were separated by reverse-phase h.p.l.c. ($C_{18}$ column) and sequenced by the manual diaminobenzoyl isothiocyanate method as previously described [31]. *N*-Terminal sequences for the α- and β-polypeptides were obtained by the same method.

## RESULTS AND DISCUSSION

### Genomic clone

The genomic clone λ JC5 contains a 13.5 kb segment of pea genomic DNA produced by partial digestion with *Eco*RI; its restriction map is given in Fig. 1(*a*). When *Eco*RI restriction digests of the genomic clone were hybridized to a labelled 'minor' legumin cDNA species, pCD40 [10], two separate hybridizing regions were detected. These fragments were subcloned into a plasmid vector (pUC9), and further restriction mapping and sequencing were carried out on them. The fragments contained gene sequences designated *LegJ* and *LegK*, located as shown in Fig. 1(*a*). Genes *LegJ* and *LegK* are arranged in tandem, in the same orientation, with *LegK* 5′ to *LegJ*. The sequences are separated by 6.5 kb (from the 3′ end of the coding sequence of gene *LegK* to the 5′ end of gene *LegJ*). The complete coding sequence of gene *LegJ* is present, but the 5′ end of the coding sequence of gene *LegK* is truncated by the end of the genomic clone.

### DNA sequences

Sequencing maps for gene *LegJ* and the available part of gene *LegK* are given in Fig. 1(*b*), and the corresponding sequences in Fig. 2. The *LegJ* sequence contains 1742 bases of coding sequence plus introns, 616 bases of 5′ flanking sequence and 611 bases of 3′ flanking sequence, whereas the *LegK* sequence runs from the end of the genomic clone, at a position corresponding to base 503 of the *LegJ* sequence, to a point 3′ to the end of the coding sequence, and thus includes 273 bases of 3′ flanking sequence.

The two genes show a high degree of homology in the coding regions (1023/1050 bases = 97%) and are thus clearly members of the same gene sub-family. The *LegJ* coding sequence predicts an open reading frame of 503 amino acid residues, containing a complete legumin subunit pair plus leader sequence [8] (22 amino acid residues of leader sequence, 300 amino acid residues of α-subunit and 181 amino acid residues of β-subunit). The start codon in gene *LegJ* has been assigned to the second ATG in the sequence after the 'TATA' box, at base 64, which is in-frame with the coding sequence. An earlier in-frame ATG is present at base 49, but the environment of this codon is not in agreement with the consensus

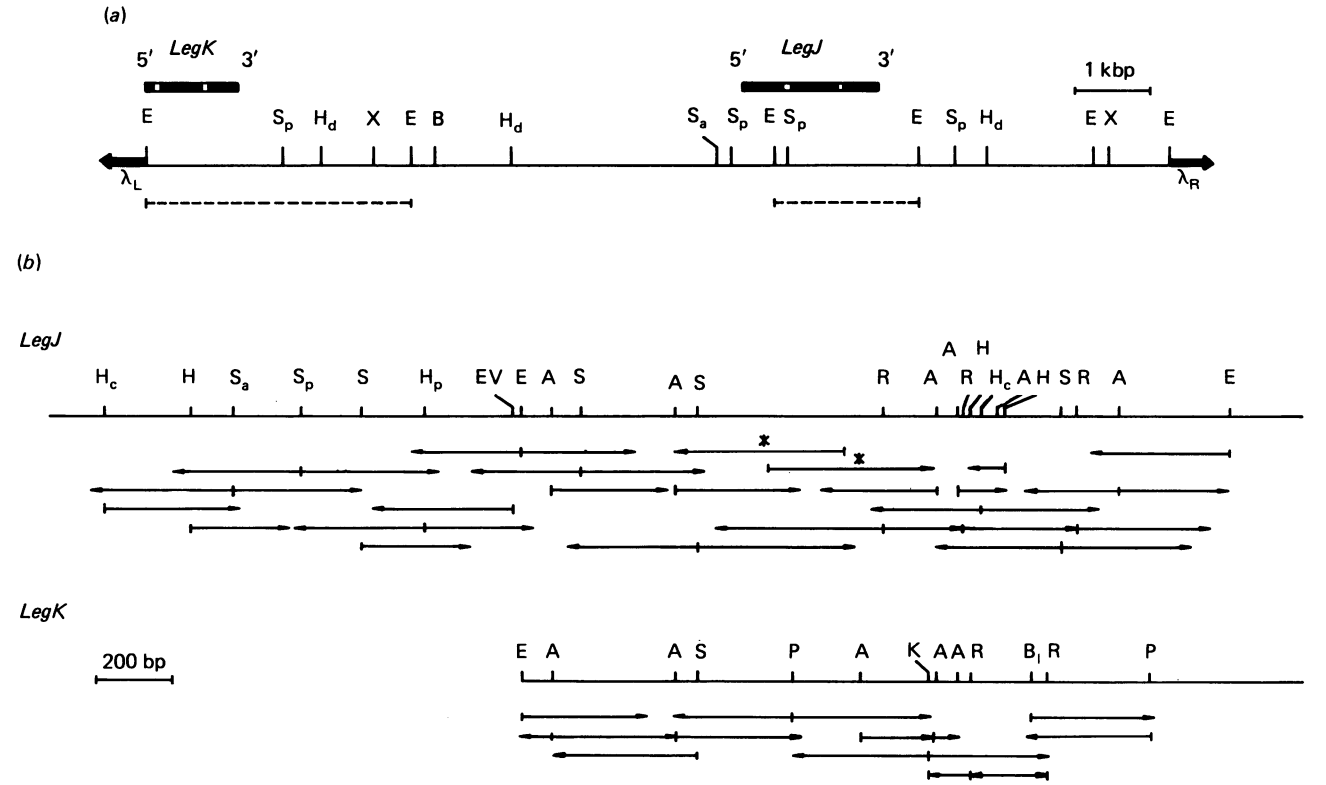



**Fig. 1. (*a*) Partial restriction map for pea genomic clone λ JC5, showing relative positions of genes *LegJ* and *LegK*, and (*b*) sequencing maps for genes *LegJ* and *LegK***

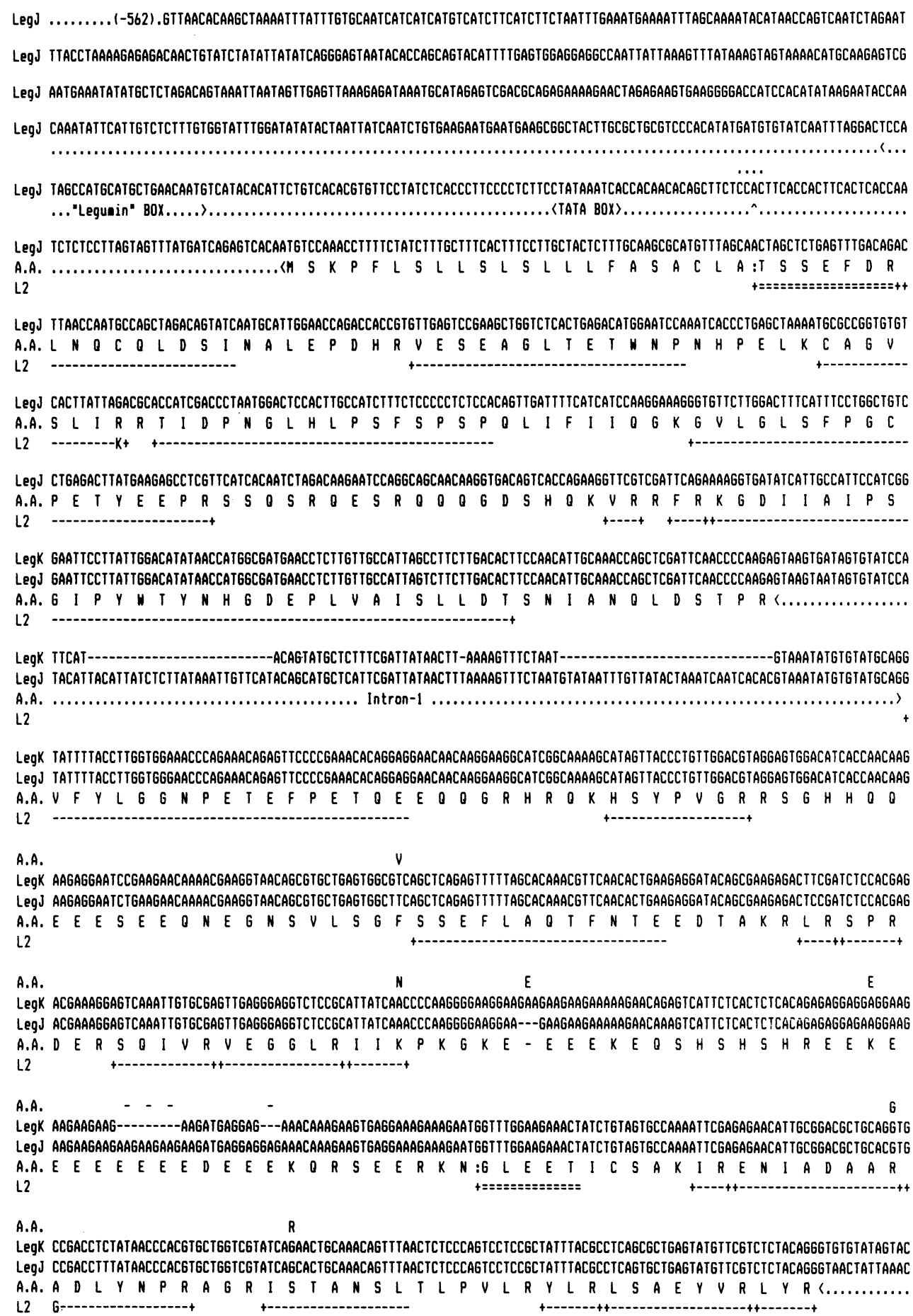
(*a*) Partial restriction map for pea genomic clone λ JC5, showing relative positions of genes *LegJ* and *LegK*. Fragments hybridizing to cDNA probe pCD40 are indicated by broken lines. (*b*) Sequencing maps for genes *LegJ* and *LegK*. Duplicated sequence runs are not indicated. Sequencing runs marked with an asterisk (*) were carried out with oligonucleotide primers. Key to restriction sites: A, *Alu*I; B, *Bam*HI; $B_1$, *Bal*I; E, *Eco*RI; EV, *Eco*RV; H, *Hae*III; $H_c$, *Hin*cII; $H_d$, *Hin*dIII; $H_p$, *Hpa*I; K, *Kpn*I; P, *Pst*I; R, *Rsa*I; S, *Sau*3AI; $S_a$, *Sal*I; $S_p$, *Sph*I; X, *Xho*I.

sequence (AACAATGGC) for plant gene translation starts [32]; 3/9 bases agree compared with 7/9 bases for the second ATG. It is therefore unlikely that this is used as a start codon. Multiple ATG sequences at the 5′ ends of genes have been noted in seed lectin genes [33]. The differences in the coding sequences between genes *LegJ* and *LegK* give a total of five in-frame codon deletions (four in gene *LegK*, one in gene *LegJ*); the remaining 27 base changes comprise 16 silent changes and 11 active changes.

The *LegJ* coding sequence predicts a polypeptide precursor of $M_r$ 57024, and a final legumin subunit pair with an α-subunit of $M_r$ 34485 and a β-subunit of $M_r$ 20300. These values are not wholly in agreement with the apparent $M_r$ values of the 'minor' legumin precursors (63000–65000; [10]) or with $M_r$ values for 'minor' legumin polypeptides (α, 38000–43000; β, 20700–21900; [28]), although data presented below suggest that the discrepancies are due to incorrect $M_r$ values being deduced from SDS/polyacrylamide-gel electrophoresis. Typical features of the legumin coding sequence are present, including cysteine residues at position 87 in the α-subunit and position 7 in the β-subunit to form the interchain disulphide bridge, an asparagine as the *C*-terminal residue of the α-subunit at which proteolysis takes place, an 'acidic' α-subunit and a 'basic' β-subunit, and a hydrophilic, charged and glutamate-rich *C*-terminal region in the α-subunit. The mature polypeptides contain no methionine residues, in contrast with 'major' legumin. Further consideration of the amino acid sequence is beyond the limits of the present paper.

The genes contain two corresponding introns, the position of which was confirmed by comparison with sequences of homologous cDNA clones (see below). The introns in gene *LegJ* are 138 and 98 bases long respectively, and those in gene *LegK* are 81 and 105 bases long. The 3′ ends of both introns in both genes show a high degree of homology (15/18 bases the same in all four), but homology at the 5′ end of intron 2 is lower. The intron boundaries are a good fit to the plant consensus sequences [34] except for the 5′ end of intron 2 in gene *LegK* (GTGTGT versus consensus GTAAGT). All boundaries obey the Breathnach–Chambon rule [35]. The overall level of homology of the introns is 56% and 71% (or 96% and 74% if deletions are ignored), which, although lower than the coding sequence homology, is still highly significant and suggests a relatively recent sequence divergence for these two genes. Like many other plant gene introns, the sequences are A+T-rich.

The 3′ flanking sequences of genes *LegJ* and *LegK* do not show significant homology over the last 60 bases of the *LegK* sequence, whereas the *LegK* 3′ flanking sequence before this region shows consistent significant homology to *LegJ* (163/217 = 75%). The two genes therefore have been sequenced far enough in the 3′



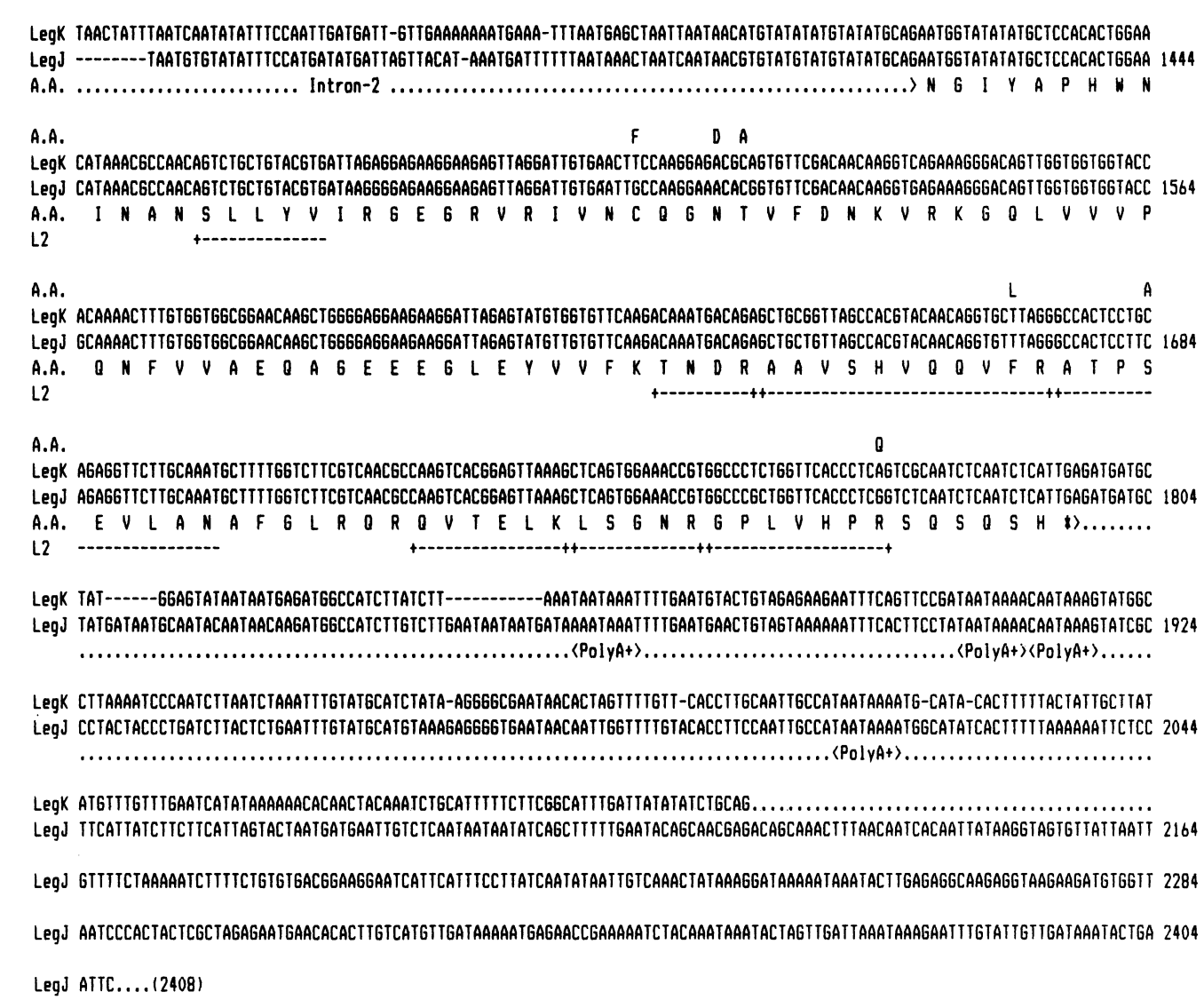



**Fig. 2. Nucleotide sequence of gene *LegJ*, and available sequence of gene *LegK***

The amino acid sequence predicted by gene *LegJ* is also shown and compared with sequences of peptides from 'minor' legumin subunit pair L2. Underlinings show extent of sequenced peptides and indicate alternative residues found in peptide sequences; alternative residues predicted by gene *LegK* are shown above the nucleotide sequence. Double underlinings indicate *N*-terminal peptides of the $\alpha$- and $\beta$-subunits of the mature protein; vertical lines indicate *N*- and *C*-termini of peptides. The transciption start shown by S1-nuclease mapping is indicated by over-dots; the base designated +1 is shown by ^. Other features are as indicated; consensus polyadenylation signals are abbreviated as PolyA+; 'Legumin' box designates the putative 5′ enhancer sequence element.

direction to have covered the region of strong homology. Both genes contain at least four potential polyadenylation sites within 220 bp of the stop codon, the first of which in gene *LegK* is of the multiple overlapping (AATAATAAA) type [8]. The second or third polyadenylation sites are the most likely to be used, since, although the sites of addition are not known in these genes, a homologous cDNA has a poly(A) tail at a point corresponding to base 1935 in the *LegJ* gene [11].

The 5′ flanking sequence of gene *LegJ* contains a clearly defined 'TATA' box, 67 bases before the start of the coding sequence, whose sequence (CCTATAAATT) is in reasonable agreement with the consensus sequence for this promoter element {T(C/G)TATA(T/A)ATA; [34]}. There is no recognizable 'CAAT' box, similarly to many other plant genes.

### Transcription start in gene *LegJ*

The start of transcription in the *LegJ* gene was determined by an S1-nuclease protection experiment. Fragments of the gene, from the *Sph*I site at position −91 to either the *Msp*I site at position 254 or the *Xba*I site at position 414, were 5′-end-labelled and hybridized to polyadenylated RNA isolated from developing pea seeds. Analysis of the DNA strand of the DNA–RNA hybrids formed after treatment with S1 nuclease on DNA sequencing gels (Fig. 3) showed protected fragments of 253–256 bases for the *Msp*I fragment and 419–421 bases for the *Xba*I fragment. The transcription start was localized to a CCAC sequence 41 bases 5′ to the start of the coding sequence, and the A residue was designated +1. The A residue is 27 bases 3′ of the TATA

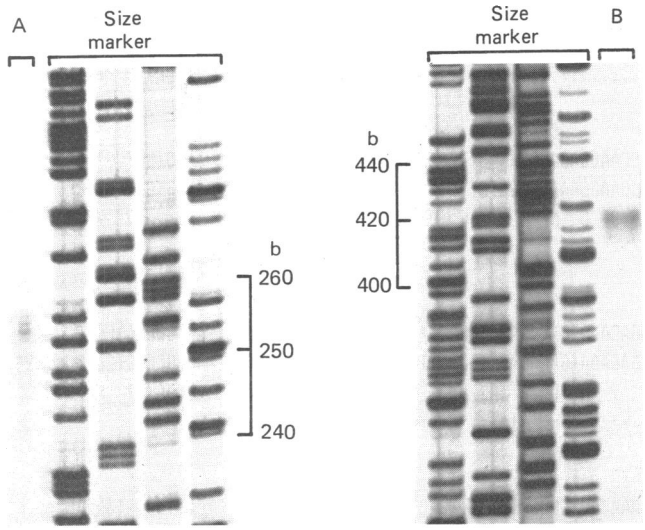



**Fig. 3. S1-nuclease mapping experiment to locate the transcription start in gene *LegJ***

The protected fragment from the *MsP*I site is run in track A, that from the *Xba*I site is run in track B. Other tracks are a size marker of known sequence, which in a separate experiment gave a fragment-size calibration exactly the same as that determined by comparison with the gene sequence.

box, in agreement with the consensus TATA–start distance [34].

### Expression of gene *LegJ*

The S1-nuclease mapping experiment described above suggested that *LegJ* was an expressed gene, since transcription of a homologous but not identical gene would not have given protected fragments of the size observed. Further evidence was obtained by hybridization of a 3′-flanking-sequence probe to total RNA isolated from developing pea seeds. Two *Eco*RI fragments from gene *LegJ*, 504–1833 (coding sequence) and 1833–2407 (3′ flanking sequence), were prepared and labelled. They were then independently hybridized to Northern blots prepared from electrophoretic separations of glyoxalated total RNA from developing pea seeds. Results are shown in Fig. 4. The 'coding-sequence' probe hybridized to a heterogeneous RNA population in the size range 2100–2400 bases, whereas the '3′-flanking-sequence' probe hybridized to a single band of size approx. 2100 bases. These results suggested that the '3′-flanking-sequence' probe was hybridizing specifically to the mRNA species corresponding to gene *LegJ*, whereas the 'coding-sequence' probe was hybridizing to mRNA species corresponding to all the expressed genes in the sub-family.

The pattern of expression of *LegJ* and homologous genes during seed development, as shown by the relative intensities of hybridization of the probes to RNA preparations isolated at different developmental stages, showed that mRNA species in this gene sub-family are present at relatively low concentrations at the early stages of cotyledon expansion, and are present at highest concentrations at the later stages of cotyledon expansion. This is a similar qualitative pattern of expression through seed development to the 'major' pea legumin genes of the *LegA* class [36], although the quantitative amounts of mRNA species present differ between 'major' and 'minor' legumin types [37].

### Homologous cDNA species to genes *LegJ* and *LegK*

Evidence that genes *LegJ* and *LegK* are both expressed genes was obtained from the sequences of two cDNA clones, which had been isolated by screening libraries prepared from polyadenylated RNA purified from developing pea seeds. One cDNA species (pLG3.121) was isolated by hybridization of the *Eco*RI fragment containing gene *LegJ* to a bank prepared from pea variety Feltham First [38]. This cDNA appeared to be a product of gene *LegJ*, since it was the same in sequence as bases 503–1456, excluding the introns. A second cDNA species, pCD40, prepared from pea variety Birte, is described elsewhere [11,31]. This cDNA species is apparently derived from gene *LegK*, implying that the gene is expressed also, and covers the region corresponding to bases 919–1934 in gene *LegJ*, excluding intron 2. The sequences are identical apart from two base substitutions between pCD40 and gene *LegK*, at bases 1700 (C in cDNA, T in gene; a silent base change in coding sequence) and 1892 (T in cDNA, C in gene; change in 3′ flanking sequence). These substitutions are most likely to represent line–line genetic variation in the *LegK* gene, since the cDNA species and the gene were obtained from different pea lines; the possibility that the cDNA species is not a product of gene *LegK* is unlikely.

### Confirmation that genes *LegJ* and *LegK* encode 'minor' legumin polypeptides

To establish the identity of the *LegJ* gene, a comparison of its encoded amino acid sequence with partial protein sequence data was made. A 'minor' legumin polypeptide pair was isolated from pea legumin; the $M_r$ values of its component polypeptides ($\alpha$, $M_r$ 43000; $\beta$, $M_r$ 21900), and its behaviour on ion-exchange chromatography, identified it as the pair designated L2 by Matta *et al.* [28]. The *N*-terminal sequences of both polypeptides were determined, and a total of 30 tryptic peptides were also sequenced. Results are presented in Fig. 2. The determined sequences totalled 242 amino acid residues, and were in agreement with the predicted sequence of gene *LegJ* except for two residues. Both G and A were found (in separate peptides) at residue 21 of the $\beta$-subunit, whereas the *LegJ* gene predicts A only, and K rather than R was indicated at residue 51 of the $\alpha$-subunit. L2 is known to exhibit microheterogeneity [28]. The amino acid sequences consistently followed the *LegJ*-predicted rather than *LegK*-predicted sequences where variant residues were determined. These results strongly suggest that polypeptide pair L2 is a product of gene *LegJ*, or a very similar gene, and prove that the gene sub-family including genes *LegJ* and *LegK* encodes polypeptides of the 'minor' legumin type.

### Sequence comparison between genes *LegJ* and *LegA*

The 'major' pea legumin gene *LegA* has been described previously by Lycett *et al.* [8]. Genes *LegA* and *LegJ* show significant homology in amino acid sequence, and in agreement with this the nucleotides of the coding sequence have an overall homology of 48 % (details not shown). Homology is stronger in some regions (e.g. the *N*-terminal half of the $\alpha$-subunit of legumin) than others (e.g. the *C*-terminal end of the $\beta$-subunit). An analysis of

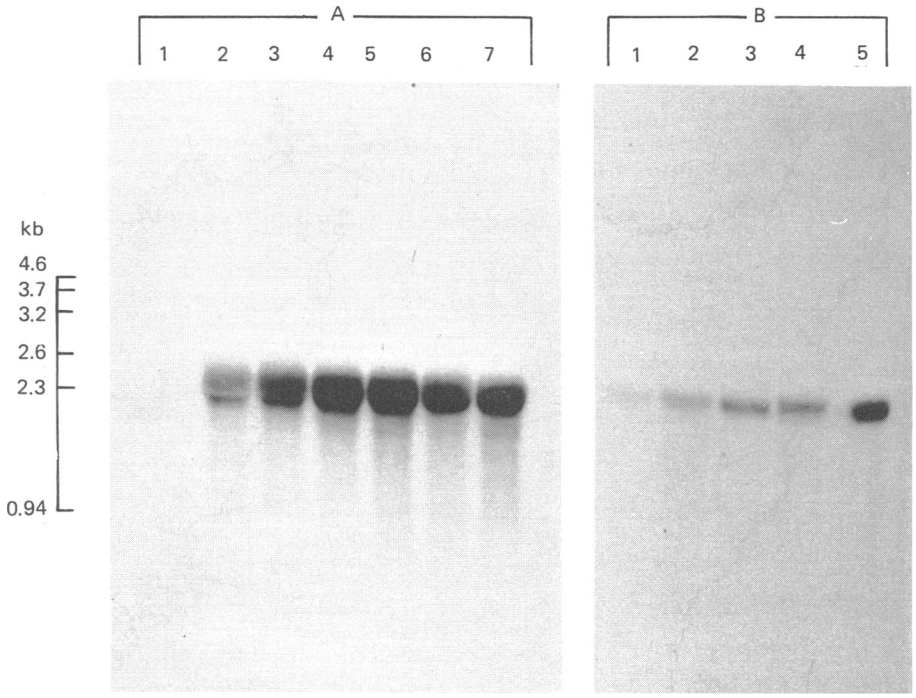



**Fig. 4. Hybridization of *LegJ* sequences to total RNA prepared from developing pea cotyledons (Northern blots)**

Tracks A1–A7, coding-sequence probe from gene *LegJ* hybridized to total RNA from cotyledons (1) 8–9 d.a.f. (days after flowering), (2) 10 d.a.f., (3) 12 d.a.f., (4) 14 d.a.f., (5) 16 d.a.f., (6) 18 d.a.f. and (7) 20 d.a.f. Tracks B1–B5, 3′-flanking-sequence probe from gene *LegJ* hybridized to total RNA from cotyledons (1) 11 d.a.f., (2) 12 d.a.f., (3) 14 d.a.f., (4) 18 d.a.f. and (5) 20 d.a.f. Cotyledon expansion occurs over the period 7–8 d.a.f. to 20–21 d.a.f. under the conditions used [22]. The molecular-size scale is taken from standard RNA species (ribosomal and cowpea-chlorotic-mottle virus RNAs) run on the original gel.

the coding sequence homology in terms of constraints on the polypeptide sequence is outside the scope of the present paper. In terms of nucleotide sequence changes, there are 60 codon deletions between the two sequences, and two single-base deletions affecting the coding frame over short regions. Two large deletions, of 21 codons in gene *LegA* relative to gene *LegJ* and 38 codons in gene *LegJ* relative to gene *LegA*, are at the *C*-terminal end of the α-subunit of legumin; the second deletion is in the sequence repeats in gene *LegA*. This region shows considerable sequence divergence, although its overall nucleotide and amino acid compositions in the two genes are similar. There is no marked divergence of the sequences of the two genes around intron 1 in gene *LegA*, which is absent from gene *LegJ*. The gaining (or losing) of intron 1 therefore seems to be a process that has taken place independently of the evolution of coding sequence. The codons that can be matched between the two genes contain an excess of active over silent base changes, although many amino acid changes are conservative. The relatively low homology between the coding sequences of genes *LegJ* and *LegA* compared with that between genes *LegJ* and *LegK* (and that between gene *LegA* and other active genes in its sub-family; S. Mahmoud & J. A. Gatehouse, unpublished work) suggests that the divergence of the two gene sub-families took place considerably earlier in evolution than the divergence of genes within sub-families. It also accounts for the absence of cross-hybridization between genes or cDNA species from different sub-families [7].

A comparison of non-coding sequences in genes *LegA* and *LegJ* shows that the 3′ flanking sequences and introns have only very weak homology, showing a greater rate of mutation in these regions. The 5′ transcribed non-translated and flanking sequences show significant homology between the two genes, homology to position −250 in gene *LegJ* being comparable with that of coding sequence. The 'TATA' box region and a region immediately 5′ to the 'CAAT' box in gene *LegA* are particularly strongly conserved, but the putative 'enhancer' sequences identified by Lycett *et al.* [8] are not strongly conserved. The strongly conserved upstream region of 28 bases (25/28 bases conserved) apparent at positions −80 to −108 in gene *LegJ* and positions −90 to −118 in gene *LegA* may be analogous to the 'enhancer' sequences found in other eukaryotic genes [39]. Homologous sequences to the entire region are also present in similar positions in other legumin genes (*Vicia faba LeB4* and soya-bean glycinin *G1*; see refs. [36] and [40]), and it seems likely that this region is involved in determining the high tissue-specific expression of these genes. Functional assays will be required to test this hypothesis. A further interesting feature in the 5′ flanking sequences is a ten-base sequence, AGGGGACCAT, present at positions −247 to −237 in gene *LegJ*, and at positions −225 to −235 in gene *LegA* on the complementary strand; this sequence is in the area where homology between the genes in the 5′ direction breaks down. It is of unknown significance.

The two genes show a marked overall similarity in nucleotide composition asymmetry, as shown by a dimensional plot (Fig. 5). Although this reflects at least

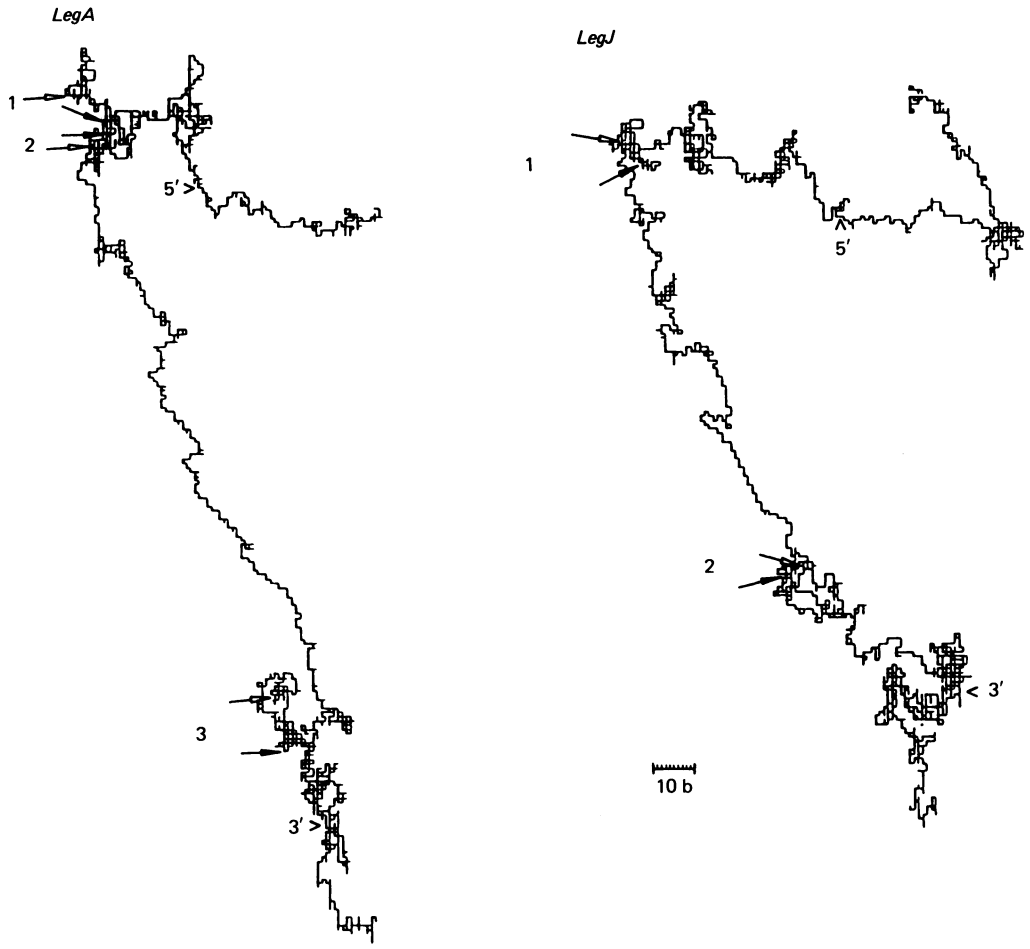



**Fig. 5. Comparison of coding and immediate flanking nucleotide sequences of genes *LegA* and *LegJ* by the 'dimensional-plot' method**

A is represented as a vector (0,−1), C as (−1,0), G as (1,0) and T as (0,1); the plot is produced by joining the end points of the resultant vectors after each base is added to the sequence. Introns are numbered as in Fig. 2 for gene *LegJ* and as in the published sequence for gene *LegA* [8], i.e. in consecutive order in a 5′ to 3′ direction. Intron splice sites are marked, the intron 5′ ends by closed-headed arrows and 3′ ends by open-headed arrows. The ends of the coding sequence (5′ and 3′) are also indicated. Note the asymmetry and similarity of sequence composition in the highly variable α-subunit *C*-terminal regions of the coding sequences (between introns 2 and 3 in gene *LegA*, 1 and 2 in gene *LegJ*).

in part the A+G-rich central sections of the genes, coding for the hydrophilic *C*-terminal ends of the α-subunits, it illustrates clearly the overall homology. The common introns in genes *LegA* and *LegJ* bound the A+G-rich central section, which may have some relevance to the concept of introns defining functional 'domains' in the protein structure.

## Comparisons with homologous sequences in other species

**'B-type' legumin gene *Vfa LeB4* from *Vicia faba* (broad bean).** Gene *LegJ* and a gene encoding a legumin precursor polypeptide from *Vicia faba*, *Vfa LeB4* [40], are homologous over a region extending from −450 bp in the 5′ flanking sequence to 265 bp 3′ to the stop codon in gene *LegJ*. The genes show a high degree of homology in their coding sequences. Homology in exons 1 and 3 is over 90%, but homology in exon 2 is markedly lower (approx. 75%); this exon contains the *C*-terminal region of the α-subunit, which shows strong variability both in terms of deletions and substitutions. This A+G-rich sequence region (see Fig. 5) is very variable in all legumin genes characterized to date, and must be concluded to be evolving in a different way to the test of the coding sequence. There is no marked preference for 'silent' base changes in any of the coding sequences.

The corresponding introns, which are of similar lengths in the pea and *Vicia faba* genes, also show a significant degree of homology, and have diverged primarily by deletions rather than substitutions. The overall homology is approx. 60%, which suggests a relatively recent divergence of the pea and *Vicia faba* genes. The 5′ flanking sequence of gene *LegJ* is approx. 80% homologous to gene *Vfa LeB4* to position −250; this homology is higher than that of the introns and shows evidence of functional constraints leading to sequence conservation. The sequences are significantly homologous to position −450, to a degree comparable with the homology of introns, giving a limit for the immediate flanking regions of the genes. There is also considerable homology in the 3′ flanking regions (74%), suggesting possible functional constraints on this sequence region.

**Soya-bean glycinin cDNA species.** Two full-length glycinin cDNA species, those encoding the A3B4 and A5A4B3 subunits [41,42], derive from genes homologous to *LegJ*, *LegK* and *Vfa LeB4*. A partial sequence of a further glycinin cDNA species and gene (that encoding the A2B1a subunits; [43]) is 'about 70%' homologous to pea gene *LegA*, but is much less homologous to genes *LegJ*, *LegK* and *Vfa LeB4* (details not shown). The division of legumin genes into sub-families that has been made in pea is thus also applicable to soya bean. Following Baumlein *et al.* [40], genes more homologous to the pea 'major' legumin genes (*LegA* class) may be designated A-type, whereas genes more homologous to the pea 'minor' legumin genes (*LegJ* class) may be designated B-type. The homology between the soya-bean A-type gene sequence (*A2B1a*) and the soya-bean B-type gene sequence is similar to that between the pea A-type and B-type gene sequences (data re-estimated from above sources). The separation of the A-type and B-type legumin gene sub-families must therefore have taken place before the divergence of any of the three species, pea, *Vicia faba* and soya bean.

The two soya-bean B-type legumin cDNA sequences are 82% homologous to each other [42], but are of significantly lower homology to genes *LegJ*, *LegK* and *Vfa LeB4* (approx. 60%; details not shown); their homologies to all three of these genes are not significantly different. There is almost complete divergence of sequences between soyabean and the other species at the *C*-terminal region of the legumin $\alpha$-subunit, confirming the extreme variability of this region noted above, although the sequence region is constrained insofar as the nucleotide composition remains A+G-rich, and the encoded amino acid residues are predominantly polar and hydrophilic. The homology of the soya-bean genes to the pea genes is significantly lower than that between the pea and *Vicia faba* genes, in agreement with the placing of pea and *Vicia faba* in the same taxonomic tribe, the Viciae, whereas soyabean belongs to the tribe Glycineae. These data allow a tentative ordering of evolutionary events: separation of A-type and B-type legumin genes, divergence of Viciae and Glycinae, separation of genes in soya-bean B-type legumin sub-family, divergence of pea and *Vicia faba*, separation of genes in pea B-type legumin sub-family.

The availability of further sequences from legumin-type genes, e.g. cruciferin from *Brassica napus*, will allow an estimate for the time of separation of the A-type and B-type legumin gene sub-families to be made, and may eventually lead to a viable molecular clock for plant species based on these genes.

We thank Dr. Claire Domoney for supplying pCD40, L. N. Gatehouse for supplying pLG3.121, Dr. R. Casey for helpful discussions, Paul Preston for technical assistance, and Professor D. Boulter for departmental facilities. Financial support from the Agricultural and Food Research Council is gratefully acknowledged.

## REFERENCES

1. Wright, D. J. & Boulter, D. (1974) Biochem. J. **141**, 413–418
2. Croy, R. R. D., Gatehouse, J. A., Evans, I. M. & Boulter, D. (1980) Planta **148**, 49–56
3. Derbyshire, E., Wright, D. J. & Boulter, D. (1976) Phytochemistry **15**, 3–24
4. Zhao, W.-M., Gatehouse, J. A. & Boulter, D. (1983) FEBS Lett. **162**, 96–102
5. Casey, R. (1979) Heredity **43**, 265–272
6. Croy, R. R. D., Lycett, G. W., Gatehouse, J. A., Yarwood, J. N. & Boulter, D. (1982) Nature (London) **295**, 76–79
7. Domoney, C. & Casey, R. (1985) Nucleic Acids Res. **13**, 687–699
8. Lycett, G. W., Croy, R. R. D., Shirsat, A. H. & Boulter, D. (1984) Nucleic Acids Res. **12**, 4493–4506
9. Bown, D., Levasseur, M., Croy, R. R. D., Boulter, D. & Gatehouse, J. A. (1985) Nucleic Acids Res. **13**, 4527–4538
10. Domoney, C. & Casey, R. (1984) Eur. J. Biochem. **139**, 321–327
11. Domoney, C., Barker, D. & Casey, R. (1986) Plant Mol. Biol. **7**, 467–474
12. Domoney, C., Ellis, T. H. N. & Davies, D. R. (1986) Mol. Gen. Genet. **202**, 280–285
13. Maniatis, T., Fritsch, E. F. & Sambrook, J. (1982) Molecular Cloning: A Laboratory Manual, Cold Spring Harbor Laboratory, Cold Spring Harbor
14. Rigby, P. W. J., Dieckmann, M., Rhodes, C. & Berg, P. (1977) J. Mol. Biol. **113**, 237–251
15. Sanger, F., Nicklen, S. & Coulson, A. R. (1977) Proc. Natl. Acad. Sci. U.S.A. **74**, 5463–5467
16. Messing, J. (1983) Methods Enzymol. **101**, 20–78
17. Biggin, M. D., Gibson, T. J. & Hong, G. F. (1983) Proc. Natl. Acad. Sci. U.S.A. **80**, 3963–3965
18. Staden, R. (1982) Nucleic Acids Res. **10**, 2951–2961
19. Gates, M. A. (1985) Nature (London) **316**, 219
20. Favaloro, I., Treisman, R. & Kamen, R. (1980) Methods Enzymol. **65**, 718–749
21. Maxam, A. M. & Gilbert, W. (1980) Methods Enzymol. **65**, 499–560
22. Gatehouse, J. A., Evans, I. M., Bown, D., Croy, R. R. D. & Boulter, D. (1982) Biochem. J. **208**, 119–127
23. Kuhn, S., Anitz, H. J. & Starlinger, P. (1979) Mol. Gen. Genet. **167**, 235–241
24. Feinberg, A. P. & Vogelstein, B. (1983) Anal. Biochem. **132**, 6–13
25. Chirgwin, J. M., Przybyla, A. E., Macdonald, R. J. & Rutter, W. J. (1979) Biochemistry **18**, 5294–5299
26. McMaster, G. K. & Carmichael, G. G. (1977) Proc. Natl. Acad. Sci. U.S.A. **74**, 4835–4838
27. Thomas, P. S. (1980) Proc. Natl. Acad. Sci. U.S.A. **77**, 5202–5205
28. Matta, N. K., Gatehouse, J. A. & Boulter, D. (1981) J. Exp. Bot. **32**, 1295–1305
29. Gatehouse, J. A., Croy, R. R. D. & Boulter, D. (1980) Biochem. J. **185**, 497–506
30. Casey R., March, J. F., Sharman, J. E. & Short, M. N. (1981) Biochim. Biophys. Acta **670**, 428–432
31. Gatehouse, J. A., Lycett, G. W., Croy, R. R. D. & Boulter, D. (1982) Biochem. J. **207**, 629–632
32. Lutke, H. A., Chow, K. C., Mickel, F. S., Moss, K. A., Kern, H. F. & Scheele, G. A. (1987) EMBO J. **6**, 43–48
33. Hoffman, L. M. & Donaldson, D. D. (1985) EMBO J. **4**, 883–889
34. Messing, J., Geraghty, D., Heidecker, G., Hu, N., Kridl, J. & Rubinstein, I. (1983) in Genetic Engineering of Plants (Kosuge, T., Meredith, C. P. & Hollaender, A., eds.), pp. 211–227, Plenum Publishing Corp., New York
35. Breathnach, R., Benoist, C., O'Hare, K., Gannon, F. & Chambon, P. (1978) Proc. Natl. Acad. Sci. U.S.A. **75**, 4853–4857
36. Gatehouse, J. A., Evans, I. M., Croy, R. R. D. & Boulter, D. (1986) Philos. Trans. R. Soc. London B **314**, 367–384
37. Domoney, C. & Casey, R. (1987) Planta **170**, 562–566
38. Gatehouse, L. N. (1986) M.Sc. Thesis, University of Durham

39. Dynan, W. S. & Tjian, R. (1985) Nature (London) **316**, 774–779
40. Baumlein, H., Wobus, U., Pustell, J. & Kafatos, F. C. (1986) Nucleic Acids Res. **14**, 2707–2720
41. Fukazawa, C., Momma, T., Hirano, H., Harada, K. & Udaka, K. (1985) J. Biol. Chem. **260**, 6234–6239
42. Momma, T., Negero, T., Hirano, H., Matsumoto, A., Udaka, K. & Fukuzawa, C. (1985) Eur. J. Biochem. **149**, 491–496
43. Marco, Y. A., Thanh, V. H., Turner, N. E., Scallon, B. J. & Nielsen, N. C. (1984) J. Biol. Chem. **259**, 13436–13441

Received 15 January 1987/24 August 1987; accepted 1 October 1987