Volume 14 Number 6 1986 Nucleic Acids Research

The legumin gene family: structure of a B type gene of *Vicia faba* and a possible legumin gene specific regulatory element

Helmut Bäumlein, Ulrich Wobus, James Pustell[1] and Fotis C.Kafatos[1,2]

Zentralinstitut für Genetik und Kulturpflanzenforschung der Akademie der Wissenschaften der DDR, DDR-4325 Gatersleben, GDR, [1]Cellular and Developmental Biology, The Biological Laboratories, Harvard University, Cambridge, MA 02138, USA and [2]Institute of Molecular Biology and Biotechnology and Department of Biology, University of Crete, Heraklio Crete 71110, Greece

Received 10 December 1985; Accepted 10 February 1986

ABSTRACT

The field bean, Vicia faba L. var. minor, possesses two sub-families of 11 S legumin genes named A and B. We isolated from a genomic library a B-type gene (LeB4) and determined its primary DNA sequence. Gene LeB4 codes for a 484 amino acid residue prepropolypeptide, encompassing a signal peptide of 22 amino acid residues, an acidic, very hydrophilic α-chain of 281 residues and a basic, somewhat hydrophobic β-chain of 181 residues. The latter two coding regions are immediately contiguous, but each is interrupted by a short intron. Type A legumin genes from soybean and pea are known to have introns in the same two positions, in addition to an extra intron (within the α-coding sequence). Sequence comparisons of legumin genes from these three plants revealed a highly conserved sequence element of at least 28 bp, centered at approximately 100 bp upstream of each cap site. The element is absent from the equivalent position of all non-legumin and other plant and fungal genes examined. We tentatively name this element "legumin box" and suggest that it may have a function in the regulation of legumin gene expression.

INTRODUCTION

Seed storage proteins of angiosperms belong to a small number of evolutionarily homologous families, one of which is the family of legumin-type proteins or 11S globulins (1). All legumin-type proteins investigated thus far consist of six nearly identical subunits (2,3). In turn, each subunit consists of an α/β pair of polypeptide chains, connected by disulfide bonds and derived from a single propolypeptide that bears the α-chain at the N-terminal side (for a recent review see 4).

We have been investigating the DNA sequences encoding legumins of the field bean, Vicia faba L. var. minor. Like other legumin-type proteins, V. faba legumins are polymorphic. Two major types of subunits, A and B, have been distinguished, initially according to the presence (A-type) or absence (B-type) of methionine, and subsequently according to amino acid analyses, sequence determinations and peptide mapping (5,6). At the DNA level (7,8), hybridization studies and cDNA sequence analyses fully confirmed that V. faba possesses a legumin multigene family, divided into A and B subfamilies.

The two legumin subfamilies are probably of more general occurrence. Methionine-rich and methionine-poor components have been characterized in soybean (9,10), and are classified as glycinin groups I and II, respectively, corresponding to the V. faba A and B subfamilies. These two groups in soybeans differ by approximately 40% at the nucleotide level (10). Comparable subfamilies may also exist in peas (see 8 for discussion). For convenience, we will use the V. faba terminology for the legumins of all species.

The characteristics of the A subfamily are known from complete DNA sequences previously determined in pea (11) and soybean (T. Sim and R. Goldberg, pers. commun.). The B subfamily is only characterized at the nucleotide level from partial cDNA sequences (10,12; Wobus, et al., in prep.) and from the recently published A_3B_4 glycinin gene (13). This paper describes the first complete genomic sequence of a gene from the B-subfamily, and defines a legumin gene-specific upstream element, possibly regulatory, that is apparently shared by the distantly homologous A- and B-subfamilies of V. faba and other legumes.

MATERIALS AND METHODS

Restriction endonucleases and T4 ligase were gifts of M. Hartmann (Jena). EcoRI was isolated according to the method of Greene et al. (14). Other batches of restriction enzymes, T4 ligase and T4 kinase were purchased from New England Biolabs, DNase I from Worthington, DNA polymerase I large fragment from Bethesda Research Laboratories, and calf intestine alkaline phosphatase grade I from Boehringer.

Vicia faba genomic DNA was isolated from leaves according to the method of Kislev and Rubinstein (15). No special precautions were taken to prevent contamination of the crude nuclear preparation by chloroplast and mitochondrial DNA.

Growth of phages, phage DNA isolation, preparation of phage arms, partial digestion of genomic DNA, size selection, ligation, preparation of packaging extracts from strains BHB2688 and BHB2690 and screening of the library were done essentially as described in Maniatis et al. (16).

Plasmid DNA was isolated according to the method of Birnboim and Doly (17). Plasmid mini preparations were performed according to Holmes and Quigley (18). For all subcloning experiments pUC18 (19) was used as a vector.

Conditions used for colony hybridization and RNA labelling by polynucleotide kinase are described in Gergen et al. (20).

For Southern blotting experiments with phage DNA and for plaque

hybridization, probes were labelled by nick translation as described in Maniatis et al. (16).

For sequence analysis the systematic strategy of Hong (21) was used. A 1.95 kb SphI fragment was cloned in both orientations in M13mp18. Linearization of the replicative form by a partial DNase I cut, cleavage at a unique restriction site in the polylinker, recircularization and transformation into JM109 generated a series of systematically deleted overlapping clones from both orientations (Fig. 1). For determination of sequences outside the SphI fragment, a 4.7 kb BglII fragment was cloned in both orientations in pUC18, and fragments from the BglII site of the polylinker to the first SphI site within the insert were prepared and subcloned in M13mp18. Sequence analyses were done according to Sanger et al. (22), using ^{35}S-dATP for labelling and buffer gradient gels to improve the resolution (23). Sequence data were processed using the computer programs of Pustell and Kafatos (24, 25, 26).

RESULTS AND DISCUSSION

Isolation of legumin B genes from a genomic library

To construct a V. faba genomic library, several methods for isolation of DNA had to be tested. The procedure of Kislev and Rubinstein (15) yielded high molecular weight, restriction enzyme digestable DNA from leaves of 4 week old plants. The genomic library was constructed by ligating 15 to 23 kb long EcoRI fragments from partial digests to Charon 4A arms followed by in vitro packaging. Approximately 2×10^6 independent recombinant phages were obtained, representing statistically the V. faba genome with 95% probability. Using characterized cDNA clones as probes (7), the amplified library was screened several times and seven phages carrying genes or gene fragments of the legumin B subfamily were selected and preliminarily characterized. Clone λVfLeB4 appeared to contain a complete gene and was chosen for detailed characterization and sequence analysis. Southern blots of various enzyme digests of that clone were hybridized with a ^{32}P-labelled poly(A)$^+$RNA preparation from middle stage cotyledons, to find fragments containing the entire coding sequence. The shortest such fragment was a 1.95 kb SphI fragment, which was contained within a 4.7 kb BglII fragment (results not shown). The SphI fragment and neighboring segments of the BglII fragment were subcloned into M13mp18 and used to construct an overlapping series of deletion clones to be used for sequencing (see Materials and Methods).

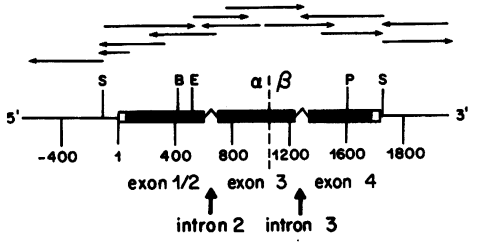

Figure 1. Diagram and sequence determination of the Vicia faba LeB4 gene. The gene is diagrammed with exons shown as rectangles (filled for coding regions, open for untranslated sequences) and with introns shown as slanted lines. The division between α and β chain coding regions is also shown (dashed line). Numbering in bp is from the putative cap site. The sequencing strategy is described in Materials and Methods. Sequences derived from various deletion clones are indicated by arrows above the diagram. Important restriction sites are indicated: S=SphI, B=BamH1, E=EcoRI, P=PstI.

Nucleotide sequences

A total of 2844 bp were determined from λVfLeB4 by the strategy depicted in Fig. 1. The sequence includes the complete gene (which is abbreviated LeB4), plus approximately 690 bp 5' flanking and approximately 330 bp 3' flanking sequences (Fig. 2). We did not sequence across the SphI sites but their uniqueness was confirmed by comparisons with homologous genes (see below) and by the recovery of predicted restriction fragments, with end-points flanking each SphI site (EcoRV and RsaI fragments for the sites 5' and 3' to the gene, respectively).

In the analysis of LeB4, valuable information was provided by comparisons with known legumin protein and cDNA sequences, plus two distantly homologous genes of the A subfamily, one from the pea, Pisum sativum (11; legA gene of clone λLeg1, abbreviated Leg1) and one from the soybean, Glycine max (T. Sim and R. Goldberg, pers. comm.; clone DA28-30, abbreviated G1). Despite their evolutionary distance, these genes share with LeB4 important landmarks, as discussed below.

The transcribed sequence of LeB4

Because of the multiplicity of B-type genes in the V. faba genome, the exact transcriptional start point is difficult to determine experimentally. However, within the SphI fragment which uniquely hybridizes with cotyledon poly(A)+ RNA (see above), and just downstream of a TATA box, we find a stretch of 16 nucleotides that differ at only 3 positions among the three genes, LeB4, Leg1 and G1 (Figure 3). That stretch includes the major 5' ends reported for

Leg1 and G1 mRNAs (11; T. Sim and R. Goldberg, pers. commun.). We propose as position +1 for LeB4 an A residue which coincides with the major 5' end of G1 mRNA, and which is conserved in all three genes as part of a TCAC element (Fig. 3). The proposed cap site resembles in location and sequence the sites described for several other plant genes (27, 28). It should be noted that the putative cap site is located within an octanucleotide, CACTTCAC, which is tandemly repeated in LeB4, once inexactly and three times exactly; only the first two repeats are evident in the Leg1 and G1 genes. This repetitiveness may be related to variations in transcriptional initiation, as observed for Leg1 (11).

The 3' end of the gene was identified by comparison with the poly(A)-terminated cDNA clone pVfc70 (Wobus et al., in prep.). That clone corresponds to a B-type gene that differs from LeB4 in the 3' untranslated region by two inserts and 27% sequence divergence in the aligned segments. However, homology at the extreme 3' end is especially high, permitting rather precise localization of the poly(A) addition site: as shown in Fig. 2, 23 nucleotides beginning with an AATAAA element and ending two nucleotides before the poly(A) of pVfc70 are identical in LeB4. This high conservation may indicate that the distal 3' untranslated sequence is significant for polyadenylation, as was shown for the 3' end of SV40 late mRNA (29, 30). Assuming that the last two nucleotides of the mRNA are mismatched between LeB4 and pVfc70, and accepting the cap site shown in Figs. 2 and 3, the predicted transcript size of LeB4 is 1829 nucleotides.

The transcribed sequence was analyzed for potential protein coding regions by a computer program (26). The analysis identified the coding strand unambiguously, and suggested (by two changes in the reading frame) the presence of two introns. Consistent with that analysis, an ATG is found in the first coding frame identified by the program, 56 nucleotides downstream from the putative cap site, and a TGA terminates the last coding frame, 126 nucleotides upstream of the poly(A) addition site. The translational start site and the first reading frame were further confirmed (see Fig. 2) by a) agreement with the consensus start site sequence proposed in (31), b) by the matching of 9 in-frame codons with the known N-terminal sequence of mature α-polypeptide chains in B legumins (A. Otto and K. Kraft, pers. comm.), and c) by the interposition of 21 additional codons between the initiator ATG and these 9 matched residues, in agreement with evidence for a ca. 1.5 kD signal peptide (32).

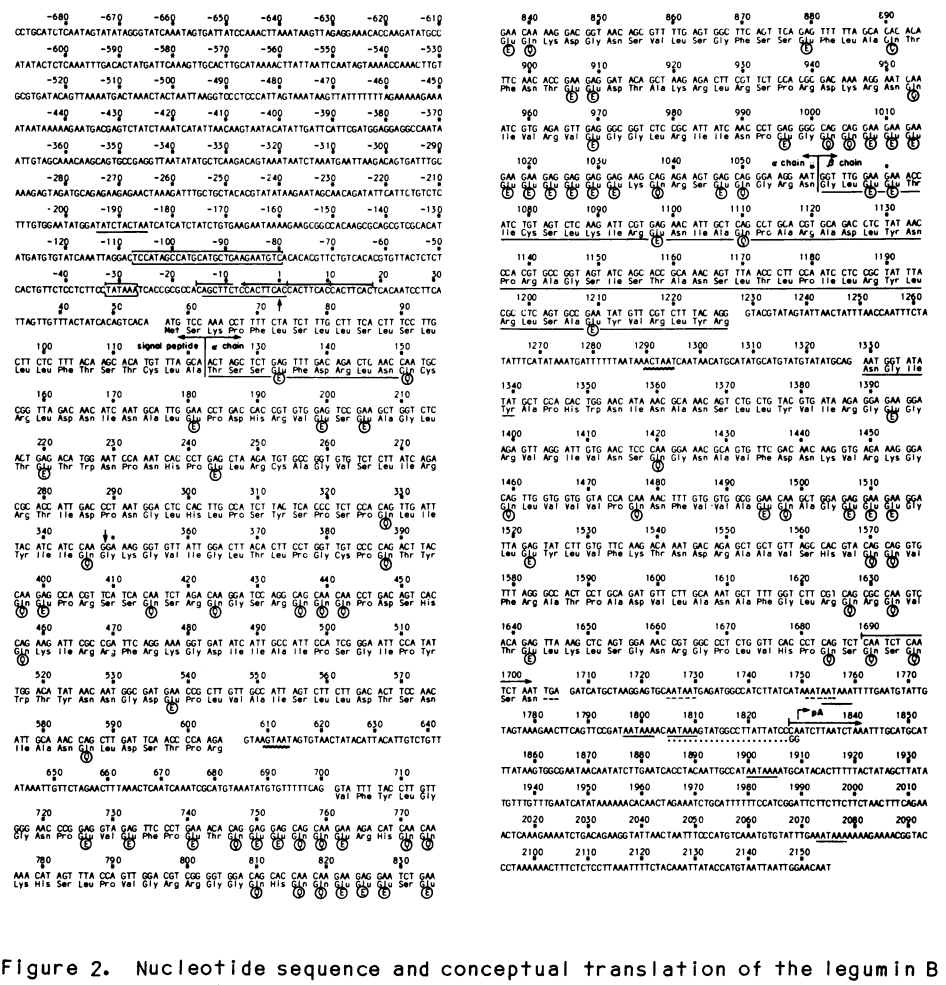

Figure 2. Nucleotide sequence and conceptual translation of the legumin B gene, LeB4. See the text for details and references.

As in Fig. 1, numbering is from the putative cap site (position 1, arrow). Underlining of positions -14 through +2 indicates a sequence block that encompasses the cap site and is conserved in legumin genes (see Fig. 3). Note the tandem repeats of CACTTCAC in the same region (overlapping of positions -15 through +18). Further upstream, boxes indicate the TATA element and the "legumin box" (cf. Fig. 3). The underlined decanucleotide centered at -190 also occurs in the G1 glycinin gene at position -480.

Coding regions are translated and assigned to the signal peptide and the αand βchains of the mature protein. The first 9 residues of the αchain and the first 60 residues of the βchain are known from protein sequencing (underlined). Note the distribution of Glu and Gln residues (circled E and Q). The arrow after position 348 indicates the location of intron 1 in other legumin genes. Within introns 2 and 3 a putative splice signal is emphasized by wavy underlining.

The multiple polyadenylation signals in the 3' untranslated region are

labelled by solid (AATAAA) and dashed (AATAAT) lines. The presumed site of poly(A) addition, inferred from the cDNA clone pVfc70, is indicated by pA. Dots between that site and the nearest upstream polyadenylation signal indicate sequence identity between pVfc70 and LeB4. The arrows at positions 1689-1702 and 1828-1841 represent a direct repeat of unknown significance.

As shown in Fig. 2, the LeB4 3' untranslated sequence is typical of plant genes, in having multiple polyadenylation signals (28, 33). Another interesting although unexplained feature is the direct repetition of a 14 nucleotide sequence (with one mismatch), almost exactly bracketing the 3' untranslated region (Fig. 2). Both the gene flanking regions and the transcribed but untranslated sequences are rich in A+T (67.5% and 67% respectively).

Intron/exon structure of LeB4

The LeB4 gene is interrupted by two short introns, 95 bp and 100 bp in length. The introns exhibit a much higher A+T content than the coding regions (76% vs. 53%), and were approximately located by the presence of termination codons as well as by changes in the apparent reading frame (see above). They were precisely positioned (Fig. 4) by comparisons with the pea and soybean genes, and in one case with the V. faba cDNA clone, pVfc70 (Wobus et al., in prep.) We designate these introns 2 and 3, since they occupy exactly the same positions as introns 2 and 3 in the A-type genes of soybean and pea. It is noteworthy that the latter two genes possess an additional intron (intron 1) that is absent from the corresponding position of the B-type gene, LeB4 (at position 348 in Fig. 2). Since Vicia and Pisum are closer taxonomically and Glycine is more distant, the presence or absence of intron 1 does not follow the phylogenetic relationships of the species; it is possible that it may be a diagnostic feature of the A- and B-subfamilies, respectively.

All splice site sequences obey the AG/GT rule and show additional homology to the donor and acceptor conserved sequences derived from plant genes (28). In addition, an internal splice signal consensus derived from mammalian and Drosophila genes (34, 35) is also matched by intron 3 and, less precisely, by intron 2 (cf. Fig. 2). After removal of the intron sequences, the predicted mRNA would measure 1634 nucleotides, in reasonable agreement with the size of legumin mRNA determined on polyacrylamide gradient gels (cf. 36).

Whereas Lycett et al. (33) observed a remarkable avoidance of CG dinucleotides in a number of analyzed plant genes, including seed storage protein genes, and specifically noted preferrential use of AGPu over CGN

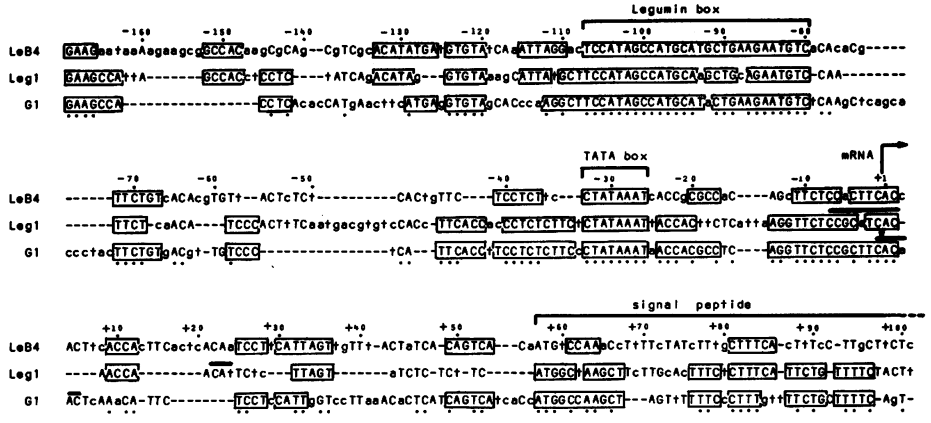

Figure 3. Sequence conservation in the 5' flanking regions and in the 5' ends of three legumin genes. Computer searches identified the major conserved elements. After manual alignment of the nucleotides between such elements, matches of four or more bases between any two genes were boxed; all matches were capitalized, while substitutions were shown in lower case and deletions were shown as dashes. Nucleotides represented in all three genes are indicated by dots below the G1 sequence. Brackets identify the signal peptide coding region, the TATA element and the highly conserved "legumin box". Heavy overlining indicates the approximate, experimentally determined 5' ends of genes Leg1 and G1. The proposed mRNA start or cap site of LeB4 (nucleotide +1) corresponds to the major 5' end of G1 (arrow). For details and references see text.

	Exon 1/2	Intron 2	Exon 3
LeB4	..CTT GAT TCA ACC CCA AGA	gtaagtaata -- 75nt -- tgttttcag	GTA TTT TAC CTT ..
Leg1	..CTT GAT CAG ATG CCT AGG	gtgagcactg -- 68nt -- tgtttgacag	AGA TTC TAT CTT ..
G1	..CTC GAC CAG ATG CCT AGG	gtgagccaca --270nt -- ttcttggcag	AGA TTC TAT CTT ..

	Exon 3	Intron 3	Exon 4
pVfc70	..TAT GTT CGT CTT TAC AGG		AAT GGT ATA TAT ..
LeB4	..TAT GTT CGT CTT TAC AGG	gtacgtatag -- 80nt -- gtatatgcag	AAT GGT ATA TAT ..
Leg1	..CAT GGA TCT CTC CAC AAA	gtatgttttt -- 80nt -- cttcatacag	AAT GCT ATG TTT ..
G1	..TTT GGA TCT CTC CGC AAG	gtacgtacat --395nt -- tacgatgcag	AAT GCA ATG TTC ..

Figure 4. Identification of the introns in LeB4. Each intron was identified by sequence comparisons with the known exon/intron structure of legumin genes Leg1 and G1. Intron 3 was also identified by its absence from the Vicia faba cDNA clone, pVfc70 which is otherwise identical with LeB4 in that vicinity. Note that the LeB4 and Leg1 introns are very similar in size, whereas those of G1 are significantly longer.

codons for arginine, in LeB4 the prevalence of CG is only moderately low: the frequency of CG is approximately half the value expected on the basis of base composition, and AGPu and CGN codons are used equally (20 occurences each).

Predicted polypeptide sequence

As defined above, the LeB4 mRNA sequence should be translated into a precursor polypeptide of 484 amino acid residues, with a calculated molecular weight of 54,430.

Bassüner et al. (32) demonstrated the co-translational cleavage of a signal peptide from the preprolegumin B of V. faba. Knowing the N-terminal amino acid sequence of the mature α-peptide (A. Otto and R. Kraft, pers. commun.) we can derive the sequence of the signal peptide from the nucleotide sequence of Fig. 2. The 22 amino acids of that peptide show typical features, such as a hydrophobic core preceded by a positively charged residue and followed by β-turn creating residues, and a signal peptidase recognition site (see 37, 38).

After signal peptide cleavage the legumin propolypeptide is known to be further processed to give rise to the disulfide bridged α/β, mature chain pair. Although the LeB4 propolypeptide has four Cys residues, there is probably only one disulphide bond formed between Cys residues 87 of the α-chain and 7 of the β-chain (Muntz and Horstmann, pers. commun.) The cleavage site leading to β-chain formation is defined by the known N-terminal amino acid sequence of the β-chain (39). Most likely that cleavage alone is responsible for α/β chain formation. It has been suggested that an upstream pair of basic residues corresponds to a second cleavage site, leading to removal of a short linker peptide from the α-chain of both glycinin (9) and Pisum legumin (40; but see also 41); however, only single basic residues are found in that region in LeB4. The maximal molecular weight of the conceptual translation product for the α-chain is 31,941, a value appreciably lower than the 36,000 estimated by SDS polyacrylamide gel electrophoresis (5). The same discrepancy was reported for the α-chain of glycinin (9). No such difference is seen for the β-chain (20,121 versus 20,000).

Although the legumin A and B sequences are recognizably homologous, they show considerable sequence divergence. Their similarities tend to be localized (Fig. 5), suggesting that certain parts of the protein structure are constrained by selection (see also 42). As in the A subfamily, the α-polypeptide of LeB4 is highly enriched in charged or very polar amino acids (58.7% Asp, Glu, Lys, Arg, His, Ser, Thr, Asn, Gln) and relatively poor in

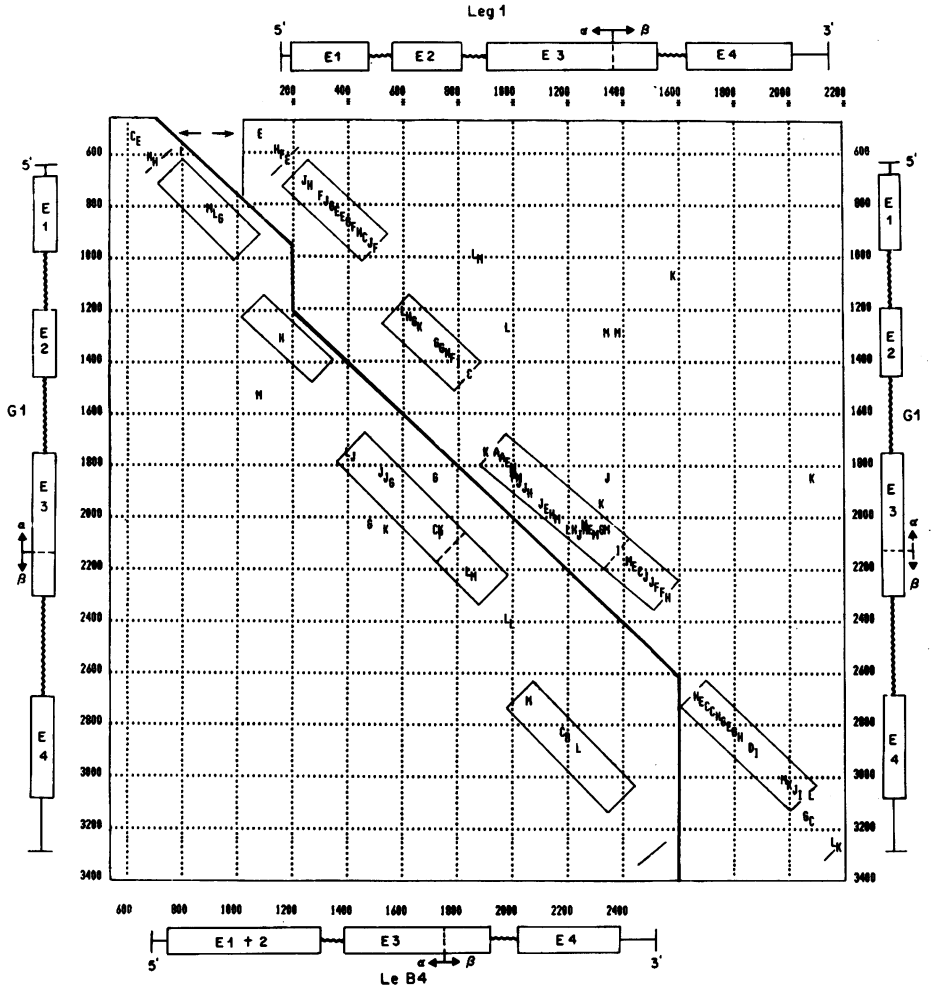

Figure 5. Matrix comparisons between legumin genes. Two matrices are shown, G1 x LeB4 (lower left) and G1 x Leg1 (upper right), separated by a zig-zag solid line. The genes are diagrammed on the axes, with coding regions of the exons shown as rectangles, introns as wavy lines, and 5' and 3' untranslated sequences as bars. The α/β chain borders are shown as dashed lines. Numbering is not from the cap site but from the 5' end of the known sequence (positions -638, -689 and -170 relative to the cap site for G1, LeB4 and Leg1, respectively). Within the matrices, the coding regions are boxed and the 5' and 3' ends are indicated with short lines. The degree of sequence homology is shown by letters in the matrices, with B = 98-99% of the maximum homology score, C = 96-97%, D = 94-95% etc. (matrix parameters: range 10, scale 0.95, hash level, jump level and step 1, compressed 20 times). Blank spaces indicate homology scores of less than 76%. Note that the 5' flanking regions (upper left corner) are substantially conserved, better than all intron

sequences (blank spaces between rectangles). For the distantly related genes, G1 vs. LeB4, the 5' flanking regions are also better conserved than the 5' and 3' untranslated sequences and almost all of the coding sequences. Arrows show matches due to the legumin box. Off-diagonal matches are largely due to Glu-rich repeats of the β-chain, most abundant in Exon 3.

non-polar amino acids (30.3% Ala, Val, Leu, Ile, Pro, Phe, Trp); Glu and Gln are the two most abundant residues (12.1% and 9.6%, respectively), especially in the C-terminal portion encoded in exon 3, where they account for 34% of the residues (see Fig. 2). Relative to the α-chain, the β-chain has fewer charged or very polar amino acids (46.4%) and more non-polar amino acids (42.5%); in hydropathy plots it appears as slightly hydrophobic, in sharp contrast to the very hydrophilic α-chain. The two chains also differ in their net charge: the α-polypeptide is acidic (16.4% acidic and 13.8% basic residues; estimated pI = 5.76), whereas the β-polypeptide is basic (8.3% acidic and 13.3% basic residues; estimated pI = 11.20).

Some but not all A-type legumins include tandem Glu-rich repeats (41). Such repeats are missing from LeB4. Other than the existence of multiple Glu- and Gln-rich clusters at several locations in the sequence, there is no obvious sign of internal repetitive structure within the gene. A more detailed comparison of A and B sequences will be presented elsewhere (Wobus et al., in prep.).

A possible legumin gene regulatory element in the 5' flanking region

The relatively distant homologies of the field bean LeB4, pea Leg1 and soybean G1 genes permitted us to search for unusually conserved sequence elements that might be of special significance.

The initial search was conducted by the matrix comparison and automatic sequence alignment programs of Pustell and Kafatos (24, 25, 26). It was immediately obvious that the proximal 5' flanking region is unusually well conserved. For example, in Fig. 5 it is evident that the 5' flanking sequences are much better conserved than the introns; in the case of the distantly related LeB4 and G1 genes, the 5' flanking sequences are also better conserved than the transcribed untranslated sequences and almost the entire coding region.

To examine the conserved sequences in more detail, the available 5' flanking sequences and the first 100 bp of the three legumin genes were aligned by computer, followed by manual alignment of the less well matched portions. The results are shown in Fig. 3.

In addition to the already noted conservation in the immediate vicinity

of the cap site, Fig. 3 shows considerable sequence homologies around the TATA motif. Short blocks of two- or three-way homology are scattered throughout the regions compared, but usually require the insertion of gaps for alignment. However, by far the strongest conservation is found in a block extending between positions -80 to -107 (for LeB4; -90 to -117 for G1 and -89 to -116 in Leg1). Here, out of 28 contiguous nucleotides, 25 are completely invariant and the rest show two-way matches. Clear cut two- and three-way matches extend further upstream for 26 additional nucleotides, requiring only a single short deletion in Leg1.

We performed a global search for the -80 to -107 element and related sequences in the plant and fungal gene subdirectory of the GenBank database (release 28.1, July 1985). The overlapping -80 to -107 (28 nt), -80 to -115 (36 nt) and -80 to -124 (45 nt) elements of LeB4 are homologous to the corresponding sequences of Leg1 and G1 to the extent of 93-96%, 86-83% and 80-82%, respectively. Even allowing for homologies as low as 65%, the consensus forms of these three overlapping elements were not found in either strand of any of the 289 entries of the subdirectory, encompassing 270 kb of DNA. The subdirectory encompassed 73 entries from higher plants, including 24 genomic clones of higher plant genes. Indeed, considering two additional published sequences not included in the database (43, 44), it could be shown that these elements are absent at the 65% level of homology from seven genes for seed storage proteins, for which 5' flanking sequences are known: five maize zein genes (clones zG99, zE19, zA1, z4 and pML1, the latter with two promoters), one wheat gliadin gene (clone pW8233) and one french bean phaseolin gene (clone 177.4). Clearly, the presence of the -80 to -124 sequence in LeB4, Leg1 and G1 is highly significant. The latter three genes are of the legumin type, whereas the seven seed storage protein genes lacking that sequence include representatives of the prolamin type (zein, gliadin) and the vicilin type (phaseolin).

Since the -80 to -107 element (and its extensions) is strongly and uniquely conserved, both in sequence and location, in the proximal 5' flanking region of all three legumin genes studied to date, we suggest that it may have a specific regulatory function, and have tentatively named it "legumin box". In comparison with other family-specific, presumptively regulatory sequence elements, such as those associated with the genes for globin (45), histone (46), chorion (47), actin (48) or heat shock proteins (49), the legumin box is unusually long and well preserved. We are currently testing its possible

regulatory significance by reverse genetics, systematically modifying the 5' flanking region of a DNA fragment that carries the LeB4 gene, transforming tobacco with that modified fragment, and testing for regulated expression of the LeB4 gene.

ACKNOWLEDGEMENTS

We are grateful to R. Goldberg for communicating the G1 sequence prior to publication. H.B. would like to thank G. Thireos and the entire Kafatos lab group at Harvard for help, and in particular N. Spoerel for advice with the sequencing method. We thank A. Czihal, T. Tölle and B. Klumpar for skillful technical assistance, and E. Fenerjian for expert secretarial work. The work was partially supported by ACS and NIH grants to F.C.K.

REFERENCES

1. Derbyshire, E., Wright, D.J. and Boulter, D. (1976) Phytochemistry 15:324.
2. Wright, D.J. and Boulter, D. (1974) Biochem. J. 141:413-318.
3. Plietz, P., Zirwer, D., Schlesier, A., Gast, K. and Damaschun, G. (1984) Biochim. Biophys. Acta 784:140-146.
4. Muntz, K., Horstmann, C. and Schlesier, B. (1985) Biol. Zbl., in press.
5. Horstmann, C. (1983) Phytochemistry 22:1861-1866.
6. Horstmann, C. (1984) Kulturpflanze (Bln). 32:S109-S116.
7. Wobus, U., Bäumlein, H., Bassüner, R., Heim, U., Jung,R., Müntz, K., Panitz, R., Saalbach, G. and Weschke, W. (1984), Kulturpflanze (Bln.), 32:S117-S120.
8. Wobus, U., Bäumlein, H., Bassüner, R., Grafe, R., Jung, R., Müntz, K., Saalbach, G. and Weschke, W. (1985) Biol. Zbl., (in press).
9. Nielsen, N.C. (1984) Phil. Trans. R. Soc. London B 304:287-296.
10. Scallon, B., Thanh, V.H., Floener, L.A. and Nielsen, N.C. (1985) 70:510-519.
11. Lycett, G.W., Croy, R.R.D., Shirsat, A.H. and Boulter, D. (1984) Nucleic Acids Res. 12:4493-4506.
12. Schuler, M.A., Ladin, B.F., Pollaco, J.C., Freyer, G. and Beachy, R.N. (1982) Nucleic Acids Res. 10:8245-8261.
13. Fukazawa, C., Momma, T., Hirano, H., Harada, K. and Ukada, K. (1985) J. Biol. Chem. 260:6234-6239.
14. Greene, P.J., Heyneker, H.L., Bolivar, F., Rodriguez, R.L., Betlach, M.C., Covarrubias, A.A., Backman, K., Russel, D.J., Tait, R. and Boyer, H. (1978) Nucleic Acids Res. 5:2373-2379.
15. Kislev, N. and Rubinstein, I. (1980) Plant Physiol. 66:1140-1143.
16. Maniatis, T., Fritsch, E.F. and Sambrook, J. (1982) Molecular Cloning. Cold Spring Harbor Laboratory, N.Y.
17. Birnboim, H.C. and Doly, J. (1979) Nucleic Acids Res. 7:1513-1517.
18. Holmes, D.S. and Quigley, J. (1981) Anal. Biochem. 114:193-196.
19. Norrander, J., Kempe, T. and Messing, J. (1983) Gene 26:101-106.
20. Gergen, J.P., Stern, R.H. and Wensink, P.C. (1979) Nucleic Acids Res. 8: 2115-2136.
21. Hong, G.F. (1982) J. Mol. Biol. 158:539-549.
22. Sanger, F., Nicklen, S. and Coulson, A.R. (1977) Proc. Natl. Acad. Sci. USA 74:5463-5467.

23. Biggin, M.D., Gibson, T.J. and Hong, G.F. (1983) Proc. Natl. Acad. Sci. USA 80:3963-3965.
24. Pustell, J. and Kafatos, F.C. (1982) Nucleic Acids Res. 10:51-59.
25. Pustell, J. and Kafatos, F.C. (1984) Nucleic Acids Res. 12:643-655.
26. Pustell, J. and Kafatos, F.C. (1986) Nucleic Acids Res. (in press).
27. Vodkin, L.O., Rhodes, P.R. and Goldberg, R.B. (1983) Cell 34:1023-1031.
28. Messing, J., Geraghty, D., Heidecker,G., Hu, N.-T., Kride, J. and Rubinstein, I. (1983) In: Kosuge, T., Meredith, C.P. and Hollaender, A (eds.), Genetic Engineering of Plants. Plenum Press, pp.211-227.
29. Sadowsky, M. and Alwine, J.C. (1984) Molec. Cell. Biol. 4:1460-1468.
30. Gil, A. and Proudfoot, N.J. (1984) Nature 312:473-474.
31. Kosak, M. (1984) Nucleic Acids Res. 12:857-872.
32. Bassüner, R., Wobus, U. and Rapoport, T. (1984) FEBS Letters 166:314-320.
33. Lycett, G.W., Delauney, A.J., Zhao, W., Gatehouse, J.A., Croy, R.R.D. and Boulter, D. (1983) FEBS Letters 153:43-46.
34. Keller, E.B. and Noon, W.A. (1984) Proc. Natl. Acad. Sci. USA 81:7417-7420.
35. Ruskin, B., Krainer, A.R., Maniatis, T. and Green, M.R. (1984) Cell 38:317-331.
36. Püchel, M., Müntz, K., Parthier, B., Aurich, O., Bassüner, R., Manteuffel, R. and Schmidt, P. (1979) Eur. J. Biochem. 96:321-329.
37. Perlman, D. and Halvorson, H.O. (1983) J. Mol. Biol. 167:391-409.
38. von Heijne, G. (1984) EMBO J. 3:2315-2318.
39. Otto, A., Kraft, R. and Etzold, G. (1984) Kulturpflanze (Bln.) 32:S219-S221.
40. Croy, R.R.D., Lycett, G.W., Gatehouse, J.A., Yarwood, J.N. and Boulter, D. (1982) Nature 295:76-79.
41. Lycett, G.W., Delauney, A.J., Zhao, W., Gatehouse, J.A., Croy, R.R.D. and Boulter, D. (1984) Plant Molecular Biology 3:91-96.
42. Argos, P., Narayana, S.V.L. and Nielsen, N.C. (1985) EMBO J. 4:1111-1117.
43. Spena, A., Viotti, A. and Pirrotta, V. (1983) J. Mol. Biol. 169:799-811.
44. Rafalski, J.A., Scheets, K., Metzler, M., Peterson, D.M., Hedgeoth, C. and Soll, D.G. (1984) EMBO J. 3:1405-1415.
45. Dierks, P., van Ooyen, A., Cochran, M.D., Dobkin, C., Reiser, J. and Weissman, C. (1983) Cell 32:695-706.
46. Grosschedl, R. and Birnstiel, M.L. (1980) Proc. Natl. Acad. Sci. USA 77:1432-1436.
47. Kafatos, F.C., Mitsialis, S.A., Spoerel, N., Mariani, B., Lingappa, J.R. and Delidakis, C. (1985) Cold Spring Harbor Symposium 50 (in press).
48. Nudel, U., Greenberg, D., Ordahl, C.P., Saxel, O., Newman, S., and Yaffe, D. (1985) Proc. Natl. Acad. Sci. USA 82:3106-3109.
49. Pelham, H.R.B. and Bienz, M. (1982) EMBO J. 1:1473-1477.