Downloaded from rstb.royalsocietypublishing.org

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

Cloning of Pea Storage Protein Genes

D. Boulter

Phil. Trans. R. Soc. Lond. B 1984 **304**, 323-332
doi: 10.1098/rstb.1984.0027

Email alerting service Receive free email alerts when new articles cite this article - sign up in the box at the top right-hand corner of the article or click **here**

To subscribe to *Phil. Trans. R. Soc. Lond. B* go to: **http://rstb.royalsocietypublishing.org/subscriptions**

This journal is © 1984 The Royal Society

Downloaded from rstb.royalsocietypublishing.org

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

Phil. Trans. R. Soc. Lond. B **304**, 323–332 (1984)
Printed in Great Britain

Cloning of pea storage protein genes

By D. Boulter

University of Durham, Department of Botany, Science Laboratories, South Road, Durham DH1 3LE

[Plate 1]

Vicilin and legumin are the major storage proteins of *Pisum sativum*. Complementary DNAs (cDNAs) have been produced from poly(A)$^+$ mRNA isolated from developing seeds and specific storage protein cDNAs cloned into pBR322. The amino acid sequences predicted from the cDNA sequences have been compared with the actual amino acid sequences derived from the purified protein subunits. These comparisons have confirmed that the legumin α and β subunits as initially synthesized are covalently joined together and that a small peptide is subsequently removed by endoproteolysis to give the disulphide linked subunits of the mature seed legumins.

Similar comparisons between the predicted amino acid sequence of vicilin cDNA clones and the amino acid sequence determined on the isolated subunits has shown that some of the 50000 M_r type subunits are subsequently cleaved to give three subunits as products, i.e. polypeptides of 19000 M_r (α), 13500 M_r (β) and 12500 M_r or 16000 M_r if glycosylated (γ). In addition to these three subunits, cleavage at one or other of the two potential cleavage sites, results in a 33000 M_r polypeptide ($\alpha+\beta$) and a 31000 M_r polypeptide tentatively identified as $\beta+\gamma$. The presence of the sequence Lys-Glu-Asn leads to cleavage on the carboxy side of Asn at the $\beta:\gamma$ cleavage site whereas the sequence Gly-Leu-Arg does not lead to cleavage. Comparable sequence data for the $\alpha:\beta$ processing site do not exist. Comparisons of the cDNA and amino acid sequence disclose the presence of a 15 or 16 amino acid residue vicilin leader sequence as well as a 12 amino acid residue C-terminal peptide which is also removed.

The codon usage of the messenger RNAs for the storage proteins are similar to those of other plant proteins and differ somewhat from animal messenger RNAs.

Complementary DNAs for specific storage proteins when used to probe different restriction enzyme digests of pea genomic DNA reveal the presence of a small number of legumin and vicilin coding sequences (two to five for legumin and three to seven for vicilin) that occur as single copies except for one vicilin sequence present in two to three copies.

Genetic mapping experiments using whole plants locate both the main legumin and the vicilin genes on chromosome 7. The main legumin subunits are coded by genes located at a single Mendelian locus *Lg-1* located on the short arm of chromosome 7 very close to the *rub* locus and the vicilin gene is located 16 map units away close to the *r* locus.

Gene libraries prepared with size fractionated partial restriction enzymic digests of pea genomic DNA ligated into both phage λ L47 and phage λ gt wes have led to the isolation of at least three similar but different legumin genomic sequences. Comparison of the λ and cDNA legumin clones suggests the presence of at least one intron in the former.

25-2

The Royal Society is collaborating with JSTOR to digitize, preserve, and extend access to *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences.* www.jstor.org

Downloaded from rstb.royalsocietypublishing.org

Introduction

Legumes in general contain two major seed storage protein types, vicilin and legumin (Derbyshire *et al.* 1976). Seeds of *Pisum sativum* (L) have significant amounts of both proteins and since a considerable body of knowledge exists about pea physiology and genetics, this species is a good choice for the study of storage protein genes. Peas are also one of the world's major legume crops.

Since the storage proteins are found only in the tissues of the developing seed (Millerd 1975) and then only in significant amounts during the middle and late stages of development, it was suspected from the onset that the genes responsible for the storage protein would belong to the class of developmentally regulated genes, i.e. those that are only switched on in specific tissues over restricted periods of time.

Amino acid and nucleotide coding sequences of storage proteins

The enzymatic conversion of poly$(A)^+$ mRNA and its insertion into bacterial plasmids for cloning is common place in molecular biology (Maniatis *et al.* 1982). Storage proteins constitute such a large proportion of the seed protein during the main phase of synthesis, that it appeared likely that the mRNAs for the storage proteins would also contribute a major share of the total cytoplasmic poly$(A)^+$ mRNA in these tissues and cloning of their cDNAs would therefore be feasible.

With the production of cDNAs from mRNA preparations from developing pea seeds (Evans *et al.* 1980) and the identification of specific cDNAs cloned into pBR322 (Croy *et al.* 1982) it was possible to compare an amino acid sequence predicted from the cDNA sequence, representing the storage protein precursor polypeptides as synthesized on the polysomes, with the actual amino acid sequence derived from the purified proteins themselves. This comparison would identify the coding sequences of the genes and allow the determination of the sequence specificity of co- and post-translational proteolysis which had already been shown to occur from cell-free translation and pulse chase experiments (Croy *et al.* 1980 and Gatehouse *et al.* 1981). The sequence characteristics of the untranslated regions of the mRNA would also be established. Nucleic acid and protein sequencing were performed as described previously (Lycett *et al.* 1983*a*; Hirano *et al.* 1982 and Boulter & Ramshaw 1975, respectively).

Construction of cDNA plasmids

The production of cDNA (Evans *et al.* 1980) and the construction of cDNA plasmids were as described previously (Croy *et al.* 1982) except that a ribonuclease inhibitor was included in later experiments and that after double strand synthesis and S1 nuclease treatment, cDNAs were size fractionated on 0.75% agarose gels and molecules of lengths of more than 1 kilobase excised from the gel and purified as described (Dretzen *et al.* 1981). These long cDNA molecules were BamHl linkered, ligated into BamHl digested pBR322 and then again size fractionated, as described above, to enrich for recombinant plasmids containing single cDNA species of lengths of more than 1 kilobase before transformation. Alternatively single stranded cDNA was prepared by using synthesized oligonucleotides as primers. These 15-mer oligonuclotides were synthesized from the 3′ to 5′ end in a series of assembly cycles involving two chemical reactions (Gait *et al.* 1982*a*) by using an apparatus made (Gait *et al.* 1982*b*) from a commercially

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES
PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

Downloaded from rstb.royalsocietypublishing.org

available kit (Omnifit Ltd, Cambridge). After antibiotic selection and identification of colonies by hybridization with nick translated ^{32}P poly(A)$^{+}$mRNA, final identification of clones was by hybrid selected translation (Croy *et al.* 1982).

Legumin

Pea legumins have molecular masses between 380000 and 410000 Da and consist of six subunit pairs, each of which is made up of a larger, relatively acidic subunit (α) and a smaller relatively basic subunit (β), joined by a disulphide bond (Gatehouse *et al.* 1980). The pairs

```
cDNA                                                  1
40 kDa    L R E Q P Q Q N E C Q L E R L N A L Q P D N R I E(S E G G F I

cDNA                                      R A T L Q R N A L R R P
40 kDa      E T W)N P N N N E F R E C G L D L L.......................
                            Q
cDNA        Y Y S N A P Q E I F I Q Q G N G Y F G M V F P G C P E T F E
40 kDa      ............................ ..........V F P G C P E T F E

cDNA        E P Q E S E Q G E G R R Y R D R H Q K V N R F R E G D I I A
40 kDa      E P Q..................... ..............................

cDNA        V P T G I V F W M Y N D Q D T P V I A V S L T D I R S S N D
40 kDa      ........................... ............................

cDNA        Q L D Q M P R R F Y L A G N H E Q E F L Q Y Q H Q Q G Q E Q
40 kDa      ..........P R P G Y L........ ......F L Q Y F H............
                        A R F

cDNA        K E N E G N N I F S G F K R D Y L E D A F N V N R H I V D R
40 kDa      ........G N N I L S G F K R S F L E D A F N V N L R I V D R

cDNA        L Q G R N E D E E K R A I V K I K G G L R I I S P P E K Q A
40 kDa      L Q G R....................... ...G G L R.................
                                                  K
cDNA        R H Q R G S P Q E E D E D E E K Q P R H Q R G S R Q E E E E
40 kDa      ............................ ...........................

cDNA        D E D E E R Q P R H Q R R R G E E E E E D K K E R R G S Q K
40 kDa      ............................ ............................

cDNA        G K S↑R R Q G D N↑G L E E T V C T A K L R L N I G P S S S P
40 kDA      .................
20 kDa                        G L E E T V C T A K L R L N I G P S S S P

cDNA        D I Y N P E A G R I K T V T S L D L P V L R W L K L S A E H
20 kDa      D L Y N P E A G R L K T V T S L D L P V L R W L K L S A E H

cDNA        G S L H K N T M F V P H Y N L N A N S I I Y A L K G R A R L
20 kDa      G S L H K N A M F V P H Y N L N A N.............K G R.....L
                    R
cDNA        Q V V N C N G N T V F A G K L E A G R A L T V P Q N Y A V A
20 kDa      Q V V N C N G N.............. ..A G R A L T V P Q N Y A V A

cDNA        A K S L N D R F S Y V A F K T N D R A G I A R L A G T S S V
20 kDa      A.........................T N D R A G I A R L A G T S S V

cDNA        I N D L P L D V V A A T F K L Q R D E A R Q L K S N N P F K
20 kDa      I N N L P L D V V A A........ ....N E A R Q L..............

cDNA        F L V P A R Q S E N R A S A
20 kDa      F L V P A R Q.........A S V
```

FIGURE 1. The predicted partially complete amino acid sequence of main legumin subunit pairs compared to partial amino acid sequences of isolated subunits. ...... amino acid sequence from Casey *et al.* (1981), —— amino acid sequence determined directly on isolated subunits by using cyanogen bromide cleaved fragments and *Staphylococcus aureus* V8 and tryptic peptides. Where two different amino acids are given for a single residue position, two identical peptides were obtained except for these residues. Sequence in parentheses by homology with soybean legumin. (N. C. Nielsen, personal communication). Arrows represent endoproteolytic processing site. Reproduced with permission from CRC Critical Review Journals (Gatehouse *et al.* 1983*a*).

Downloaded from rstb.royalsocietypublishing.org

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

may be divided into three types, the main type (conventional legumin subunits) and two minor types, designated big and small respectively (Matta *et al.* 1981). Cell-free translation of poly(A)$^+$ mRNA preparations from developing pea seeds has demonstrated that the subunit pairs are synthesized covalently joined together and that they are subsequently cleaved to yield the two subunits (Croy *et al.* 1980).

A predicted partially complete amino acid sequence for 'main' legumin subunit pairs is given in figure 1, and is compared to the partial amino acid sequence for the legumin α and β subunits determined by direct analysis of the polypeptides. The cDNA sequence does not extend to the N-terminus of the legumin subunit pair which has been determined by Casey *et al.* (1981). The data show that the predicted amino acid sequence continues unbroken upstream from the N-terminal sequences of the β subunit into the C-terminal sequences of the α subunit without start or stop codons in the same reading frame. Just before the N-terminus of the β subunit, a sequence of two basic residues, Arg-Arg, is found. By analogy with the proteolytic processing of several animal protein precursors, for example proinsulins and the corticotropin–lipotropin precursor (Cohen *et al.* 1980), it is likely that this sequence is the cleavage site of precursor legumin and that further processing of the linking sequence occurs since the β subunit has Gly at its N-terminus. The site is in a polar region of the polypeptide and resembles that found at the β:γ cleavage site of vicilin (see below) in that acidic/amide amino acids predominate and the cleavage is on the C-terminal side of an asparagine residue (Arg-Arg-Glu-Asp-Asn/Gly-Lys-Glu-Glu). Present cDNA clones are not long enough to show whether legumin contains a 'leader' sequence.

The three legumin cDNA clones sequenced so far contain a single amber stop codon. Examination of their 3′ untranslated regions show that in contrast to the vicilin cDNAs (see below) they have the sequence AATAAATAAA 19 bases from the polyadenylation site, i.e. two overlapping repeats of the archetypal sequence. Zein (Geraghty *et al.* 1981) and leghaemoglobin C (Hyldig-Nielsen *et al.* 1982), also contain the overlapping repeat sequence, but not in the expected position relative to the polyadenylation sequence; GATAAA and AATAAGAAA are found in the expected positions however. The latter sequence variant is also found in the 3′ untranslated region of legumin cDNA, but not in the expected position 15–30 nucleotides from the polyadenylation site (Lycett *et al.* 1983*b*). Furthermore the legumin cDNAs unlike the vicilin cDNAs (see below) contain the sequence ATTTCAGTGC just to the 5′ side of the polyadenylation site similar to the conserved sequence often found straddling or on the 3′ side of the polyadenylation site in animal genomic sequences (Benoist *et al.* 1980).

Vicilin

Vicilin, M_r about 150000, is composed of three 50000 type subunits, some of which are 'nicked' so that under dissociating conditions they separate on gels (Gatehouse *et al.* 1981). Depending on gel conditions, as many as four different components can be seen on one dimensional gels with M_r between 47000 and 51000. Cell-free translation of poly(A)$^+$ mRNA isolated from cotyledons at various stages of seed development show that the initial translational products on the ribosomes are 50000 M_r type subunits only (Evans *et al.* 1979). In contrast, vicilin isolated from seeds and separated into constituent subunits on sodium dodecyl sulphate (SDS) gels, displays several subunits (Gatehouse *et al.* 1981).

A predicted complete amino acid sequence for vicilin is given in figure 2 and is compared to the partial or complete amino acid sequences of the different vicilin subunits (Hirano *et al.*

Downloaded from rstb.royalsocietypublishing.org

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

```
pDUB7    M L L A I A F L A S V C V S S R S D Q E N P F I F K S N R F Q T L Y E N E N
19 kDa                                 R S D Q E..........K S N R F Q T L F Q N E N
                                     G F
33 kDa                                †R S D Q E N P F I F K S N R F Q T L Y E N E N
                                                                       F
50 kDa                                     S D Q E N P F I F K.......F Q T L Y E N E N

pDUB7    G H I R L L Q K F D K R S K I F E N L Q N Y R L L E Y K S K P H T L F L P Q
19 kDa   G H I R L L Q K F D Q R S K I F E N L Q N Y R L L E Y K S K P H.......L P Q
33 kDa   G H I R L L Q K F D Q R S K I F E N L Q N Y R L L E Y K S K P R.......L P N
             V       K   D
50 kDa   G H I R....... .............I F E N L Q N Y R...........................

pDUB7    Y T D A D F I L V V L S G K A T L T V L K S N D R N S F N L E R G D A I K L
19 kDa   Q N D A D............S G K A T L T V L K.........N S F N L E R G D A I K L
33 kDa   Y.....................S G K A T L T V L K S N D R N S F N L E R G D A I K L
                                                         Y
50 kDa   ...........................A F L T V L L P N D R N S F N L E R..........L

pDUB7    P A G S I A Y F A N R D D N E E P R V L D L A I P V N K P G Q L Q S F L
19 kDa   P A G T I A...L A N R D D N E D L R V L D L A I P V N R P G Q L Q......
33 kDa   P A G T I A Y L A N R D D N E D L R V L D L A I P V N K P G Q L Q Q F L
                                                          R     N
50 kDa   P A G S I A Y.........D D N E.......V L D L A I P V N R P G Q..........

pDUB2                                                    †E I E K I L L E E H
pDUB7    L S G T Q N Q K S S L S G F S K N I L E A A F N T N Y E       V     Q Q
19 kDa   ...............................N I L E A A F N T N Y E.......K V L L E....
                                                 S
33 kDa   L L G T Q N Q.........S G F S K N I L E A A F N T N Y E E I E K V L L E Q H
           S                                     S
50 kDa   ...............................N I L E A S F N T D............V L L E Q..

pDUB2    E K E T H H R R G L R D K R Q Q S Q E K N V
pDUB7      Q   P Q       S   K   R     E I N   E
19 kDa   ........          †D R R Q E L S N E N V
33 kDa   E Q E P Q H G R G L R D R R Q Q S Q E K N V
               T H   R
50 kDa   ............... ..........................

pDUB2    I V K V S K K Q I E E L S K N A K S S S K K S V S S R S E P F N L K S S D P I
pDUB7            R D                                 E G         R   R N

13.5 kDa I V K V S R R Q I E E L S K N A K S S S R K S V S S E S G P F N L R S E D P I
                     E                                 N
33 kDa   ......V S R E Q I E E L S K N A K S S S R R S V S S E S G P F N L R S S D P I
                                                                           R N
50 kDa   ............................... ............S V S S K S E P F N L K S S D P I
                                                       E

pDUB2    Y S N Q Y G K F F E I T P K K N P Q L Q D L D I F V N Y V E I K E G S L W L P
pDUB7        K F               E   Q                     S   D     V       L
13.5 kDa Y S N N S G K F F E I T P E K N Q Q I Q D I D I F V N S V D I K E G S L L I P
33 kDa   Y S N N Y G K F F E I T P E K N Q Q L Q D L D I F V N S D.......E G S L L L P
50 kDa   Y S.......... ...............................................E G S L L L P

pDUB2    H Y N S R A I V I V T V N E G K G D F E L V G Q R N E N Q Q G L R E E D D E E E E Q
pDUB7    N                                                     - - - G K   N   K
13.5 kDa N Y N S R A I L V I V*V N E G K G D F E L V G Q R N E N Q - - - G K E N†D H E E E Q
33 kDa   N Y N S R..................G K G D F E L V G Q R N E N Q - - - G K E N†
50 kDa   H Y N S R......................................N E N Q Q G L R.................

pDUB2    R E E E T K N Q V Q S Y K A K L T P G D V F V I P A G H P V A V R A S S N L N
pDUB7    -         S K       L   R     S                         I N       D
16 kDa   - E E E T S K Q V Q L Y R A K L S P G D V F V I P A G H P V A I   A S S D L N
50 kDa   ............................... .............................. R A S S N L D

pDUB2    L L G F G I N A E N N Q R N F L A G E E D
pDUB4/7    I                     E
16 kDa   L I G F G I N A E N N E R N F L A G E E D
50 kDa   L I G F G I N A..........N F L A G D....

pDUB2    N V I S Q I Q I P V K D L A F P G S A Q E V D R L L E N Q K Q S Y F A N A Q P
pDUB4              V E R                   S H
16 kDa   N V I S Q V E R P V K E L A F P G S S H E V D R L L E N Q K Q S Y F A N A Q P
50 kDa   .....................D L A F P G S S Q E V D R.............................

pDUB2    Q Q R E T R S Q E I K E H L Y S I L G A F
16 kDa   L Q R E T R S Q E†
50 kDa   ......E T R S Q E
```

FIGURE 2. The predicted complete amino acid sequence for vicilia compared with partial or complete amino acid sequences of the isolated polypeptide subunits. Amino acid sequence data were determined directly on the isolated vicilin subunits (Hirano *et al.* 1982; Lycett *et al.* 1983*a* and Gatehouse *et al.* 1983*b*). Where more than one amino acid was detected in the protein sequence these are shown one above the other. Arrows: N or C terminal amino acids were determined directly; asterisk: several amino acids omitted to allow lining up for maximum homology; --- indicates spaces inserted to allow lining up for maximum homology; ..... sequence not determined. The amino acid sequences predicted by the cDNA clones are shown for comparison. Three clones were used, pDUB7, pDUB2 and pDUB4, all of which show considerable sequence homology and overlap. Where two cDNA sequences overlap, the sequence of pDUB2 is shown in full and the sequence of pDUB4/7 is shown only where it differs from that of pDUB2. Reproduced with permission from IRL Press Ltd.

Downloaded from rstb.royalsocietypublishing.org

BIOLOGICAL SCIENCES THE ROYAL SOCIETY PHILOSOPHICAL TRANSACTIONS OF

1982; Lycett *et al.* 1983*a*). The vicilin subunits of M_r less than approximately 50000 are very similar in sequence to parts of the 50000 M_r precursor and their derivation from the 50000 M_r subunit is shown in figure 3 (Gatehouse *et al.* 1982). The generalized vicilin precursor polypeptide has two sites at which post-translational proteolysis may occur. Cleavage at both these sites gives three 'subunits' as products, polypeptides of 19000 M_r (α), 13500 M_r (β) and

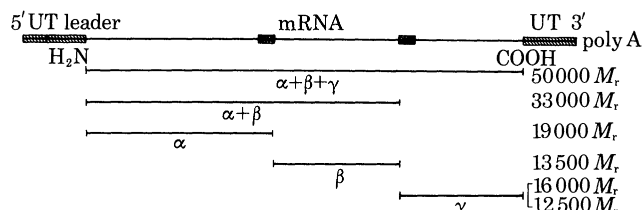


FIGURE 3. Derivation of vicilin peptides from the precursors of the 50000 M_r type subunits. The top line is a diagrammatic representation of the vicilin messenger RNA.

12500 M_r or 16000 M_r if glycosylated (γ). If neither of the potential cleavage sites undergoes proteolysis, the resulting intact 50000 M_r polypeptide is denoted ($\alpha+\beta+\gamma$). Finally, cleavage at one potential processing site but not the other leads to two possibilities: cleavage at β:γ site gives ($\alpha+\beta$), which is the 33000 M_r polypeptide, and γ, while cleavage at the α:β site gives α and ($\beta+\gamma$) which has been tentatively identified as a minor vicilin polypeptide of 31000 M_r.

The C-terminus of the β and the N-terminus of γ subunits are asparagine and aspartic residues respectively, thereby locating accurately the β:γ site of cleavage. The 33000 M_t subunit has asparagine at its C-terminus confirming the position of the cleavage site. Both the β and the 33000 M_r subunit ($\alpha+\beta$) have a tripeptide sequence 'missing' just before this site as compared with the amino acid sequence of the 50000 subunit. Vicilin cDNA clone pDUB2 has a nucleotide sequence whose predicted amino acid sequence is the same as the determined amino acid sequence of the 50000 M_r subunit in this region. Another vicilin clone pDUB7 is similar in sequence except in the area of this site. Normally the 50000 M_r subunit is not processed and a comparison of the sequence of pDUB2 that hybrid selects 50000 M_r subunit and pDUB7 that hybrid selects 47000 M_r subunit which is cleaved at one or both sites, should indicate the sequence specificity for cleavage. The data are given in figure 4 and suggest that if the sequence Lys-Glu-Asn is present the polypeptide is cleaved on the carboxy terminal side of Asn whereas the sequence Gly-Leu-Arg does not lead to cleavage.

Suitable comparable data do not exist for the α:β processing site. All the clones so far sequenced have Gly-Leu-Arg or Ser-Leu-Lys at the potential cleavage site and it is suspected that their products are unprocessed at this site especially as Gly-Leu-Arg signals non-cleavage when present at the β:γ site.

A comparison of the 5′ of vicilin cDNAs with the N-termini of 50000, 33000 and 19000 M_r polypeptides indicates that there is a leader sequence that is removed, since the peptides start with R or GF, yet there is a reading frame upstream of this residue equivalent to between 15 or 16 amino acid residues before a 'start' Met codon is reached (figure 2). Both Higgins & Spencer (1981) and Gatehouse *et al.* (1981) have presented evidence for the removal of leader sequences from vicilin in accordance with the signal hypothesis of Blobel & Dobberstein (1975). It is possible that there is another Met codon further upstream and the signal peptide is longer,

Downloaded from rstb.royalsocietypublishing.org

but evidence from cell-free translation suggests a leader sequence of molecular mass 1000–2000 (Higgins & Spencer 1981).

Comparison between the C-terminus of the 16000 M_r polypeptide with the 3′ region of the nucleotide coding sequence of vicilin cDNA suggests that a 12 amino acid residue peptide is removed post-translationally (see figure 2).

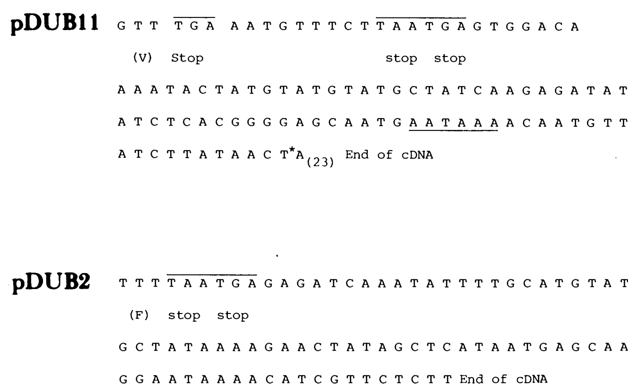


FIGURE 4. The scheme represents 3′ untranslated region of vicilin cDNA clones. Amino acid in parentheses has been predicted from cDNA sequence.

All the vicilin clones showed a simple pattern in their untranslated 3′regions of a single AATAAA sequence as do many animal clones, but unlike animals no TTTTCAGTCG sequence was found. In pDUB11 for example the putative poly(A) signal sequence, AATAAA is followed 20 nucleotides downstream by a poly(A) sequence (figure 4*a*).

In different cDNA clones the functional stop codon is either the first of a tandem pair of terminator codons (see figure 4*b*) or a stop codon (overlined) situated four codons upstream from this pair as in pDUB11 (figure 4*a*).

Codon usage

The mRNAs from animals are known to have a characteristic codon usage pattern. When this aspect of the cDNAs for legumin and vicilin were examined (Lycett *et al.* 1983*b*) it was clear that the codon usage in their sequences was different to that of animals. Codons containing the nucleotide CG were little used, e.g. the arginine codons AGA and AGG were preferred over the other four. Generally, animal systems tend to avoid codons ending in AA, AU, UA or UU (Grantham *et al.* 1981), but this was not observed with the plant cDNAs (Lycett *et al.* 1983*b*).

Genomic coding sequences

Complementary DNAs for specific storage proteins can be used to probe different restriction enzyme digest of genomic DNA to detect the number of different genomic sequences complementary to the coding sequences, and their copy number (Young *et al.* 1981). Results from these experiments show the presence of a small number of legumin and vicilin coding sequences, two to five for legumin, three to seven for vicilin (figure 5, plate 1 and Gatehouse *et al.* 1983*b*). These were present in single copies, except that one vicilin sequence was present

Downloaded from rstb.royalsocietypublishing.org

in two to three copies (Gatehouse *et al.* 1983*b*). Although there were small differences in the numbers of fragments detected when using different cDNA probes for the same storage protein, due in some cases to the presence of a restriction site in the gene, the general conclusion that there are a few gene sequences for the storage proteins remains the same. The pattern was the same for leaves and cotyledons and whether the cDNA probes were synthesized by enzymic copying of poly$(A)^+$ mRNA or constructed by using synthetic oligonucleotide primers (Gait *et al.* 1982*a*, *b*).

Variations in subunit band patterns analysed by gel electrophoresis have been used by Hynes (1968) and Thomson & Schroeder (1978) to suggest that some storage protein genes behave as single Mendelian genes.

Published (Matta & Gatehouse 1982) and unpublished work from this laboratory locates both the main legumin and vicilin genes on chromosome 7; the convicilin gene is on chromosome 2. Although minor legumin components (big and small legumin) occur, over 90% of the legumin subunits (main legumin) were coded for by genes located at this single Mendelian locus. As expected the acidic and basic subunits did not segregate independently. The legumin locus, *Lg-1*, is located on the short arm of chromosome 7 very close to the *rub* locus. Davies (1980) had shown previously that the legumin gene was linked in the *r* locus on chromosome 7. The vicilin locus is 16 map units away close to the *r* locus. It also behaves as a single locus in genetic crosses (J. A. Gatehouse and S. A. Mahmoud, unpublished observations). The storage protein genes, therefore, do not constitute an operon. A crossover unit in classical genetics is approx. 6×10^6 base pairs for a chromosome of the size of chromosome 7: thus, even though the several vicilin and legumin genes identified by hybridization probes each act as single Mendelian genes in crosses there could be considerable spacer regions between the different genes of either locus, and genomic clones are required to elucidate the extent of possible clustering.

Gene libraries have been prepared with size fractionated (about 20 kilobases in length) partial restriction enzyme digests of *Pisum* genomic DNA ligated into phage λ L47 and into phage λ gt wes by using standard methods (Maniatis *et al.* 1982). Plaques have been selected using ^{32}P nick-translated legumin cDNA inserts and 'minipreps' made. At least three different legumin genes with similar but different restriction patterns have been isolated. Legumin cDNA and mRNA binds strongly to appropriate restriction fragments of these cloned genes.

Examination of various smaller fragments cut out from the λ legumin clones shows them to be larger by 100–200 base pairs than the corresponding fragments cut from the legumin cDNA, this suggests the presence of at least one intron in the former.

Discussion

The storage protein genes of *Pisum* are an example of a phenomenon now quite commonly found in eukaryotic organisms, whereby specialized cells produce large amounts of specific proteins over a short time period by regulating the expression of a family of closely related genes. Separate gene families exist for legumin and vicilin, each of which consists of relatively few similar genes, although in the case of vicilin these genes may not all cross-hybridize with each other under stringent conditions.

As well as cleavage of N and C terminal peptides, endoproteolysis may also occur. Comparison of vicilin cDNA sequences with those of phaseolin (Sun *et al.* 1981) and conglycinin

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES
PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

Downloaded from rstb.royalsocietypublishing.org

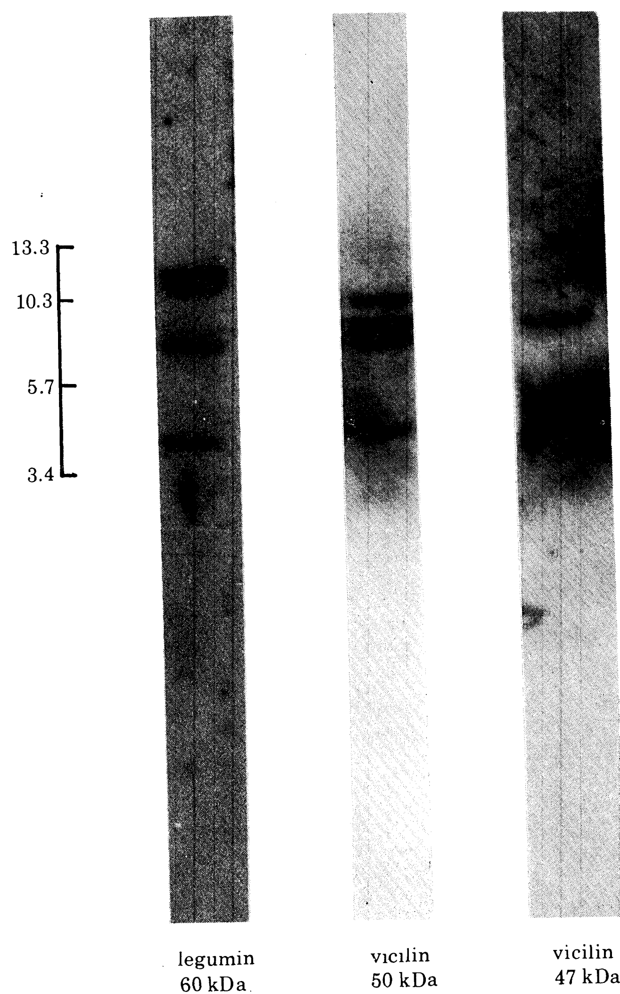


FIGURE 5. Detection of fragments encoding legumin and vicilin polypeptides in EcoRI digests of pea genomic DNA. Scale is in kilobases. Digested DNA was analysed by agarose gel electrophoresis, transferred to nitrocellulose paper, and probed with nick-translated pDUB1 (legumin) pDUB2 (vicilin 50 kDa) and pDUB4 (vicilin 47 kDa). Reproduced with permission from CRC Critical Review Journals.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

Downloaded from rstb.royalsocietypublishing.org

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

α subunit (Schuler *et al.* 1982*a*, *b*) the 7 *S* proteins of *Phaseolus* and *Glycine* respectively, shows evolutionary divergence around the post-translational endoproteolytic processing sites. These sites are in hydrophilic regions presumably on the outside of the molecules, whereas sequences coding for regions within the pea vicilin smaller subunits (α, β, γ etc) are relatively conserved in the different genera (Lycett *et al.* 1983*a*).

Considerable variation has been noted in the polyadenylation site sequence signals in plant genes, i.e. presence of a single AATAAA sequence or an overlapping double archetypal sequence or the presence of alternative sequences such as GATAAA (Lycett *et al.* 1983*b*). This is again reflected in the fact that pea vicilin contains a single AATAAA sequence whereas conglycinin of soyabean has a double overlapping sequence.

I gratefully acknowledge the help of my colleagues Dr R. R. D. Croy, Dr M. I. Evans, Dr J. A. Gatehouse, Dr G. W. Lycett, Mr A. Delauney, Mr S. H. Mahmoud and Mr A. Shirsat, and for permission to present some of their unpublished results.

References

Benoist, C., O'Hare, K. Breathnach, R. & Chambon, P. 1980 The ovalbumin gene-sequence of putative control regions. *Nucl. Acids Res.* **8**, 127–143.

Blobel, G. & Dobberstein, B. 1975 Transfer of proteins across membranes. I. Presence of proteolytically processed and unprocessed mascent immunoglobulin light chains on membrane-bound ribosomes of murine mydoma. *J. Cell Biol.* **67**, 835–851.

Boulter, D. & Ramshaw, J. A. M. 1975 Amino acid sequence analysis of proteins. In *The chemistry and biochemistry of plant proteins* (Ed. J. B. Harborne & C. F. van Summere), pp. 1–30. London: Academic Press.

Casey, R., March, J. F., Sharman, J. E. & Short, M. N. 1981 The purification, N-terminal amino acid sequence and some other properties of an α^{M}-subunit of legumin from the pea (*Pisum sativum* L.) *Biochim. biophys. Acta* **670**, 428–432.

Cohen, S. N., Chang, A. C. Y., Nakanishi, S., Inouse, A., Kita, T., Nakamwa, M. & Numa, S. 1980 Studies of cloned DNA encoding the structure for the bovine cortisotropin-beta-hyoprotein precursor protein. *Ann. N.Y. Acad. Sci.* **343**, 415–424.

Croy, R. R. D., Gatehouse, J. A., Evans, I. M. & Boulter, D. 1980 Characterization of the storage protein subunits synthesized *in vitro* by polyribosomes and RNA from developing pea (*Pisum sativum* L.) I. Legumin. *Planta* **148**, 49–56.

Croy, R. R. D., Lycett, G. W. Gatehouse, J. A., Yarwood, J. N. & Boulter, D. 1982 Cloning and analysis of cDNAs encoding plant storage protein precursors. *Nature, Lond.* **295**, 76–79.

Davies, D. R. 1980 The *r* locus and legumin synthesis in *Pisum sativum Biochem. Genet.* **18**, 1207–1219.

Derbyshire, E., Wright, D. J. & Boulter, D. 1976 Legumin and vicilin, storage proteins of legume seeds. *Phytochemistry* **15**, 3–24.

Dretzen, G., Bellard, M., Sassone-Corsi, P. & Chambon, P. 1981 A reliable method for the recovery of DNA fragments from agarose and acrylamide gels. *Analyt. Biochem.* **112**, 295–298.

Evans, I. M., Croy, R. R. D., Hutchinson, P., Boulter, D., Payne, P. I. & Gordon, M. E. 1979 Cell free synthesis of some storage protein subunits by polyribosomes and RNA from developing seeds of pea (*Pisum sativum* L.) *Planta* **144**, 455–462.

Evans, I. M., Croy, R. R. D., Brown, P. & Boulter, D. 1980 Synthesis of complementary DNAs to partially purified mRNAs coding for the storage proteins of *Pisum sativum. Biochim. biophys. Acta* **610**, 81–95.

Gait, M. J., Mathes, H. W. D., Singh, M., Sproat, B. S. & Titmas, R. C. 1982*a* Rapid synthesis of oligodexyribonucleotides VII. Solid phase synthesis of oligodeoxyribonucleotides by a continuous flow phosphotrieter method on a Kieselguhr-polyamide support. *Nucl. Acids Res.* **10**, 6243–6254.

Gait, M. J., Mathes, H. W. D., Singh, M., Sproat, B. S. & Titmas, R. C. 1982*b* In Chemical and enzymatic gene synthesis – a laboratory manual (ed. H. G. Gassen & A. Lang). Weinheim: Verlag Chemie. (In the press.)

Gatehouse, J. A., Croy, R. R. D. & Boulter, D. 1980 Isoelectric-focusing properties and carbohydrate content of pea (*Pisum sativum*) legumin. *Biochem. J.* **185**, 497–503.

Gatehouse, J. A., Croy, R. R. D., Morton, H., Tyler, M. & Boulter, D. 1981 Characterization and subunit structures of the vicilin storage proteins of pea (*Pisum sativum* L.) *Eur. J. Biochem.* **118**, 627–633.

Gatehouse, J. A., Lycett, G. W., Croy, R. R. D. & Boulter, D. 1982 The post-translational proteolysis of the subunits of vicilin from pea (*Pisum sativum* L.) *Biochem. J.* **207**, 629–632.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

Downloaded from rstb.royalsocietypublishing.org

Gatehouse, J. A., Croy, R. R. D. & Boulter, D. 1983*a* Synthesis and structure of pea seed storage proteins. *Crit. Revs Pl. Sci.* (In the press.)

Gatehouse, J. A., Lycett, G. W., Delauney, A. J., Croy, R. R. D. & Boulter, D. 1983*b* Sequence specificity of the post-translational proteolytic cleavage of vicilin, a seed storage protein of pea (*Pisum sativum* L.) *Biochem. J.* **212**, 427–432.

Geraghty, D., Peifer, M. A., Rubenstein, I. & Messing, J. 1981 The primary structure of a plant storage protein: zein. *Nucl. Acids Res.* **9**, 5163–5175.

Grantham, R., Gautier, C., Jacobzone, M. & Mercier, R. 1981 Codon catalog usage is a genome strategy modulated for gene expressivity. *Nucl. Acids Res.* **9**, 43–73.

Higgins, T. J. V. & Spencer, D. 1981 Precursor forms of pea vicilin subunits *Pl. Physiol.* **67**, 205–211.

Hirano, H., Gatehouse, J. A. & Boulter, D. 1982 The complete amino acid sequence of a subunit of the vicilin seed storage protein of Pea (*Pisum sativum* L.) *FEBS Lett.* **145**, 99–102.

Hyldig-Nielson, J. J., Jensen, E. Ø., Paludan, K., Wiborg, O., Garret, R., Jorgensen, P. & Marcker, K. A. 1982 The structures of two leghaemoglobin genes from soybean. *Nucl. Acids Res.* **10**, 689–701.

Hynes, M. J. 1968 Genetically controlled variants of a storage protein in *Pisum sativum*. *Aust. J. biol. Sci.* **21**, 827–829.

Lycett, G. W., Delauney, A. J., Gatehouse, J. A., Gilroy, J., Croy, R. R. D. & Boulter, D. 1983*a* The vicilin gene family of pea (*Pisum sativum* L.). A complete coding sequence for preprovicilin. *Nucl. Acids Res.* **11**, 2367–2380.

Lycett, G. W., Delauney, A. J. & Croy, R. R. D. 1983*b* Are plant genes different? *FEBS Lett.* **153**, 43–46.

Maniatis, T., Fritsch, E. F. & Samcrook, J. 1982 *Molecular cloning. A laboratory manual.* pp. 212–246. Cold Spring Harbor Laboratory.

Matta, N. K., Gatehouse, J. A. & Boulter, D. 1981 Molecular and subunit heterogeneity of legumin of *Pisum sativum* (garden pea) a multidimensional gel electrophoretic study. *J. exp. Bot.* **32**, 1295–1307.

Matta, N. K. & Gatehouse, J. A. 1982 Inheritance and mapping of storage protein genes in *Pisum sativum* L. *Heredity* **48**(3), 383–392.

Millerd, A. 1975 Biochemistry of legume seed proteins. *A. Rev. Pl. Physiol.* **26**, 53–72.

Schuler, M. A., Schmitt, E. S. & Beachy, R. N. 1982*a* Closely related families of genes code for the α and α′ subunits of the soybean 7 *S* storage protein complex. *Nucl. Acids Res.* **10**, 8225–8244.

Schuler, M. A., Ladin, B. F., Polacco, J. C., Freyer, G. & Beachy, R. N. 1982*b* Structural sequences are conserved in the genes coding for the α, α′ and β-subunits of the soybean 7*S* seed storage protein. *Nucl. Acids Res.* **10**, 8245–8261.

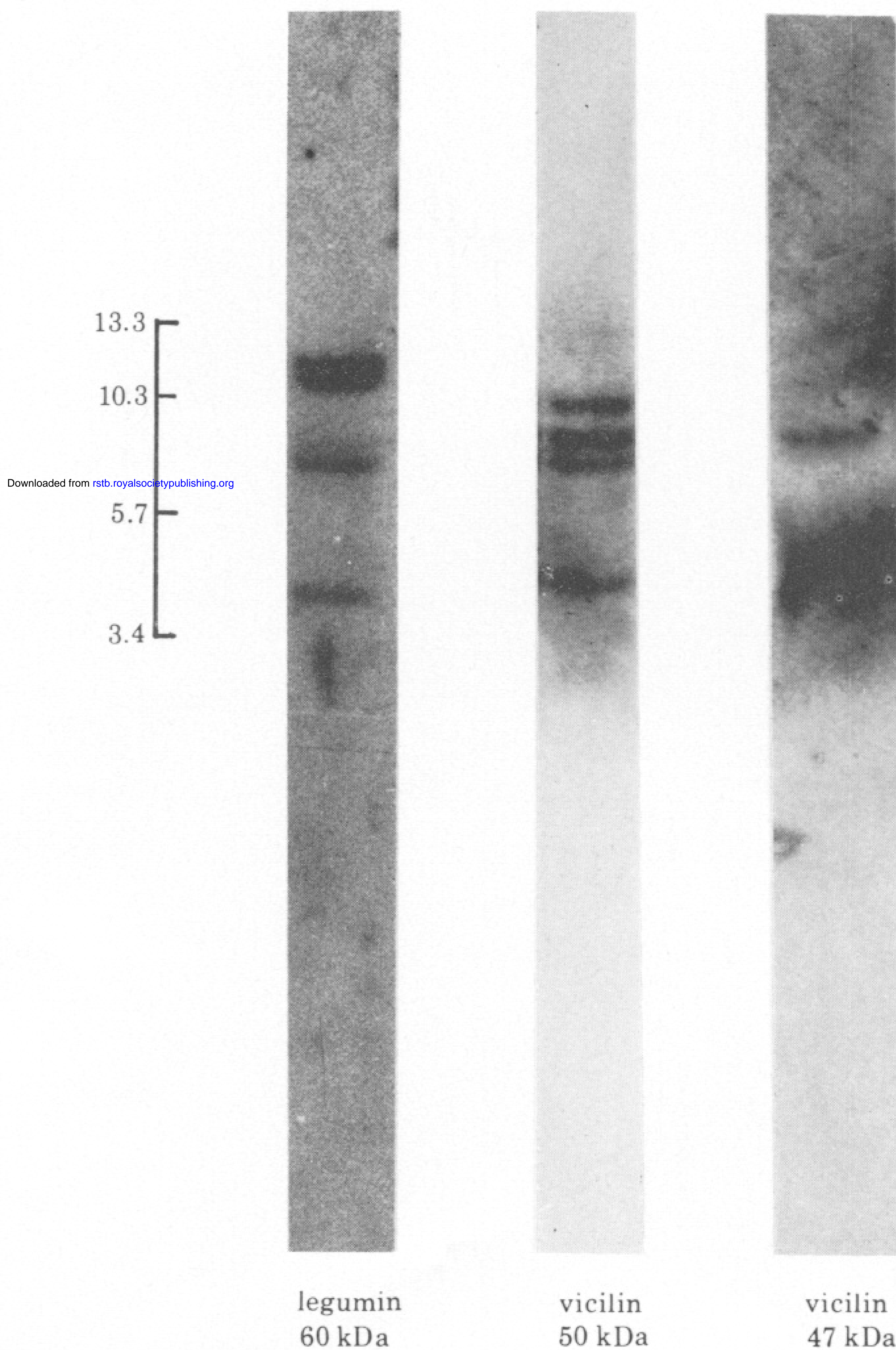
Sun, M., Slightom, J. L. & Hall, T. C. 1981 Intervening sequences in a plant gene – comparison of the partial sequence of cDNA and genomic DNA of French bean phaseolin. *Nature, Lond.* **289**, 37.

Thompson, J. A. & Schroeder, H. E. 1978 Cotyledonary storage proteins in *Pisum sativum*. II. Hereditary variation in components of the legumin and vicilin fractions. *Aust. J. Pl. Physol.* **5**, 281–294.

Young, R. A., Hagenbüchle, O. & Schibler, U. 1981 A single mouse α-amylase gene specifies two different tissue-specific mRNAs. *Cell* **23**, 451–458.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES



Downloaded from rstb.royalsocietypublishing.org

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

FIGURE 5. Detection of fragments encoding legumin and vicilin polypeptides in EcoRI digests of pea genomic DNA. Scale is in kilobases. Digested DNA was analysed by agarose gel electrophoresis, transferred to nitrocellulose paper, and probed with nick-translated pDUB1 (legumin) pDUB2 (vicilin 50 kDa) and pDUB4 (vicilin 47 kDa). Reproduced with permission from CRC Critical Review Journals.