The sequence of a gene encoding convicilin from pea (*Pisum sativum* L.) shows that convicilin differs from vicilin by an insertion near the *N*-terminus

David BOWN,* T. H. Noel ELLIS† and John A. GATEHOUSE*‡

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

The sequence of a gene encoding convicilin, a seed storage protein in pea (*Pisum sativum* L.), is reported. This gene, designated *cvcA*, is one of a sub-family of two active genes. The transcription start of *cvcA* was mapped. Convicilin genes are expressed in developing pea seed cotyledons, with maximum levels of the corresponding mRNA species present at 16–18 days after flowering. The gene sequence shows that convicilin is similar to vicilin, but differs by the insertion of a 121-amino-acid sequence near the *N*-terminus of the protein. This inserted sequence is very hydrophilic and has a high proportion of charged and acidic residues; it is of a similar amino acid composition to the sequences found near the *C*-terminal of the α -subunit in pea legumin genes, but is not directly homologous with them. Comparison of this sequence with the ‘inserted’ sequence in soya-bean (*Glycine max*) conglycinin (a homologous vicilin-type protein) suggests that the two insertions were independent events. The 5′ flanking sequence of the gene contains several putative regulatory elements, besides a consensus promoter sequence.

INTRODUCTION

Convicilin has been termed a ‘third storage protein’ in pea seeds, in addition to legumin and vicilin [1]. It can be purified from both legumin and vicilin, and it consists solely of polypeptides of M_r approx. 71000. It does not thus contain polypeptides found in either of the two major storage proteins [2]. On the other hand, convicilin is antigenically similar to vicilin [1], and it is possible to produce molecules containing both vicilin and convicilin polypeptides; for this reason, some authors have considered that convicilin and vicilin are the same protein [3]. Sequence data for a partial cDNA clone, pCD 59, identified as encoding convicilin by hybrid-release translation, supported this view, since the deduced amino acid sequence was strongly homologous with that of vicilin [4,5]. However, pCD 59 did not hybridize to vicilin cDNA species [5] or vicilin genes [6].

Variation in the mobility of convicilin polypeptides, on SDS/polyacrylamide-gel electrophoresis, between pea lines has allowed a convicilin locus, designated ‘*cvc*’, to be mapped to chromosome 2 in pea [7]; it is distinct from any vicilin locus so far identified [8,9]. Convicilin has been shown to be encoded by a small gene family; hybridization of the cDNA clones pCD 59 and pCD 75 (a longer version of pCD 59; [5]) to genomic DNA restricted with endonucleases detected one or two hybridizing fragments, depending on which probe was used [5,6,9].

The isolation of a genomic clone containing a convicilin gene, putatively corresponding to the *cvc* locus, has been described [9]. The present paper reports the sequence of this gene and its flanking regions, and shows that convicilin genes in pea (*Pisum sativum* L.) form a sub-family of the total family of vicilin-type genes.

MATERIALS AND METHODS

Materials

Pea seeds of the cultivar (cv.) Feltham First were obtained from Suttons Seeds, Torquay, Devon, U.K.; seeds of cv. Dark Skinned Perfection were from S. Dobie and Son, Torquay, Devon, U.K. The isolation of the genomic clone lambda JC4, and its sub-clone pJC 4-100, from a genomic library prepared from DNA isolated from *Pisum sativum* cv. Dark Skinned Perfection has been described previously [9]. Reagents and enzymes for M13 DNA sequencing were from Gibco/BRL (Gibco, Paisley, Renfrewshire, Scotland, U.K.); restriction enzymes were supplied by Northumbrian Biologicals, Cramlington, Northd., U.K. S1 nuclease and other enzymes were from BCL, Lewes, East Sussex, U.K. Radiochemicals were supplied by Amersham International, Amersham, Bucks., U.K. Other reagents used were of analytical quality wherever possible. Nitrocellulose filters were type BA85 (Schleicher und Schuell) from Anderman and Co., East Molesey, Surrey, U.K.

Methods

DNA sequencing. Restriction mapping on pJC 4-100 was carried out by conventional methods [10]. Preparation of subclones from pJC 4-100 in pUC18 or pUC19, preparation of sequencing subclones in M13 mp18 or mp19, preparation of single-stranded DNA, and dideoxynucleotide DNA sequencing using [α -³⁵S]thio-dATP were also carried out by standard techniques [11–14]. The sequence given was determined by overlapping sequences from subclones; both strands of the DNA were fully sequenced. Sequences were analysed by diagonal dot-matrix comparisons [15], using a

These sequence data have been submitted to the EMBL/GenBank Data Libraries under the accession number Y00721.

‡ To whom correspondence and reprint requests should be addressed.

program written by ourselves and by manual comparisons supplemented by sequence-handling software (programs NNALN and FASTP, kindly supplied by Dr. W. Pearson). Hydrophilicity profiles were plotted using the method of Hopp & Wood [16].

Blotting techniques. Restriction fragments from pJC 4-100 or its subclones were isolated from low-gelling-temperature agarose gels [17] and labelled with [α - 32 P]dCTP (400 Ci/mmol; 100 μ Ci used/0.2–0.5 μ g of DNA) by nick translation [18]. 'Southern' blots of agarose-gel separations of restriction fragments, or digests of pea leaf genomic DNA (purified as in [19]) with restriction enzymes, were prepared and hybridized to denatured labelled probes in $5 \times$ SSC ($1 \times$ SSC is 0.15 M-NaCl/0.015 M-sodium citrate buffer, pH 7.2)/ $2 \times$ Denhardt's solution ($1 \times$ Denhardt's solution is 0.02% Ficoll/0.02% bovine serum albumin/0.02% polyvinylpyrrolidone)/denatured herring sperm DNA (100 μ g/ml), at 65 °C as described in [20]; subsequent washes were to a hybridization stringency of $0.1 \times$ SSC at 65 °C. 'Northern' blots of agarose-gel separations of glyoxalated total RNA samples (prepared from pea (cv. Feltham First) cotyledons at different developmental stages as previously described [21] were prepared and hybridized to denatured labelled probes in $5 \times$ SSC, $2 \times$ Denhardt's solution/denatured herring sperm DNA (200 μ g/ml)/50% (v/v) formamide, at 42 °C [22]; subsequent washes were to a hybridization stringency of $0.1 \times$ SSC/0.1% SDS at 50 °C. Densitometry of autoradiographs, obtained by exposing the washed blots to preflashed X-ray film at –80 °C, was carried out on an LKB (Bromma, Sweden) Ultrascan XL densitometer.

S1 mapping. S1 mapping was carried out as described by Favaloro *et al.* [23]. Each assay mixture contained 5 μ g of polyadenylated RNA, prepared from pea (cv. Feltham First) cotyledons at a mid-development stage (14–15 days after flowering) as previously described [24], and at least 0.2 μ g (approx. 2×10^6 c.p.m.) of DNA probe, 5' end-labelled [25] with [γ - 32 P]-ATP (6000 Ci/mmol; 50 μ Ci used/0.2–0.5 μ g of DNA). The protected fragment after S1 digestion was run on a DNA sequencing gel, and its 3' end was mapped by running a DNA sequencing reaction that covered the same region of sequence on the same strand, and had been primed by an oligonucleotide primer whose 5' end corresponded to the site of labelling, in adjacent tracks. Controls omitting RNA were carried out.

Protein sequencing. Convicilin was purified as previously described [1]. Portions (2 mg) of the protein, dissolved in 0.1% trifluoroacetic acid, were subjected to h.p.l.c. (Vydac reverse-phase C₁₈ column; elution with a gradient of acetonitrile in 0.1% trifluoroacetic acid) to remove traces of vicilin. Convicilin polypeptides were digested with trypsin, and the resulting peptides were separated by h.p.l.c. and sequenced by the manual dansylbenzoyl isothiocyanate method, as previously described [26]. N-Terminal sequences for convicilin were obtained by automated sequence determination on an Applied Biosystems model 371A protein sequencer, with online h.p.l.c. residue identification. A 0.3 mg sample of protein was used per determination.

RESULTS

Genomic clone

A partial restriction map for the genomic subclone pJC 4-100 has been published previously [9]. A revised and detailed map, showing the position of the gene and the region sequenced, is given in Fig. 1(a). The clone contains approx. 8 kb of sequence 5' flanking the convicilin coding sequence, and approx. 3 kb of 3' flanking sequence; these regions do not contain sequences hybridizing to probes from the *cvc* coding sequence [9; results not shown]. Regions of this clone outside the sequenced region are not discussed further in the present paper.

The convicilin gene

The sequencing map for the convicilin gene is given in Fig. 1(b), and the complete sequence of the gene and its immediate 3' and 5' flanking regions is given in Fig. 2. We have designated this gene '*cvcA*'. The predicted sequence of the encoded protein was deduced by homology with vicilin and by the presence of an open reading frame at the 5' end, and is also shown in Fig. 2. The coding nucleotide sequence is interrupted by five introns, whose positions could be inferred from the predicted and determined protein sequence (the present paper) and from the nucleotide sequences of the convicilin cDNA species pCD 59 [5], the homologous *Phaseolus vulgaris* (French bean) vicilin (phaseolin) gene [27] and homologous pea vicilin cDNA species and genes ([28,29]; J. A. Gatehouse, D. Bown, M. Levasseur, R. Sawyer & T. H. N. Ellis, unpublished work). The sequence from start codon to stop codon thus contains six exons, of 661, 176, 75, 324, 283 and 197 bases respectively, and five introns, of 151, 103, 103, 88 and 97 bases respectively. The encoded amino acid sequence is 571 amino acids in length, and predicts a precursor polypeptide of M_r 66986; when the leader sequence of 28 amino acids (see below) is subtracted the predicted M_r for the mature polypeptide is 63928. The discrepancy between this value and the polypeptide M_r determined for convicilin (71000) is discussed below.

The 3' flanking sequence given extends for 428 bases after the stop codon; a further 450 bases of sequence have been determined, but do not show any significant features and will not be discussed further. Two polyadenylation sites are present in the 3' flanking sequence, 119 and 134 bases after the stop codon; the first of these is of the multiple overlapping type (AATAAATAAA) often found in plant genes [30]. The 5' flanking sequence contains a good match to the consensus sequence for a plant gene 'TATA' box [31] 66 bases before the start codon (CTATAAATA). Other sequence features in this region are discussed below.

Partial sequence of convicilin

The identity of the gene *cvcA* was confirmed by comparing its predicted protein sequence with partial protein sequence data from convicilin. In all, 16 residues at the N-terminus of convicilin and an additional 75 residues from 14 tryptic peptides were determined. Results are shown in Fig. 2. The determined sequences agree fully with the sequence predicted by *cvcA* and show that the first 28 residues of the predicted sequence are not present in the mature polypeptide. These removed residues constitute a typical 'leader' sequence [32]. At

Fig. 1. Restriction map of the clone pJC4-100 containing *cvcA*, and sequencing map of *cvcA*

Key to restriction site symbols on sequencing map; B, *Bgl*II; D, *Dra*I; E, *Eco*RV; H, *Hind*III; Hp, *Hpa*I; K, *Kpn*I (= *Asp*718I; N, *Nsi*I; S, *Sst*I; Sp, *Ssp*I; X, *Xba*I.

amino acid 209, two residues were found in tryptic peptides; N, as predicted by *cvcA*, and Q (one-letter notation). Peptides were obtained from all six exons, showing that the assignment of intron positions was valid.

Expression of *cvcA*

An S1 mapping experiment was carried out to confirm the expression of *cvcA* and to locate the transcription start. The *Asp*718I restriction fragment, covering bases -561 to 143 in *cvcA*, was isolated and 5'-end-labelled. After hybridization of the labelled fragment to polyadenylated RNA isolated from developing pea cotyledons, the nucleic acids were treated with S1 nuclease and analysed by gel electrophoresis. Results are shown in Fig. 3(a). Protected fragments of 139–150 bases were obtained, suggesting that an mRNA had identical sequence with the probe from base 143 in *cvcA* to a region 24–35 bases 5' to the ATG start codon. The base designated '+1' was that giving the most intense band in the S1 mapping assay, i.e. the underlined base in the protected sequence region, CATCATCTAAAG. Protected fragments extending to the A bases in the consensus transcription start sequences -CATC- [31] in the above region were observed, but gave less intense bands in the S1 mapping assay. Control experiments

with no RNA present gave no protected fragment. A further S1 mapping experiment, with the *Nsi*I–*Eco*RV restriction fragment, covering bases -382 to 257 in *cvcA*, gave protected fragments ending in the region -8 to +2. In this case both the S1 mapping assay and its control with no RNA present gave protected fragments corresponding in length to the original probe.

The developmental expression of convicilin genes was also studied by hybridization of part of the sequence of this gene to total RNA prepared from pea cotyledons at different stages of seed development. The probe fragment was chosen to include only the 5'-end of the coding sequence of the gene to avoid cross-hybridization to vicilin mRNA species. Pea cotyledon RNA was glyoxalated, size-fractionated by electrophoresis and blotted on to nitrocellulose before hybridization to the *Sst*I–*Bgl*II (bases -176 to 462) fragment of *cvcA*, labelled by nick translation. The results of this experiment are shown in Fig. 3(b). The probe hybridized to two bands of similar mobility on the Northern blot, corresponding to mRNA species of approx. 2650 and 2500 bases; the larger of the two species consistently gave a more intense hybridization signal, the ratio of the integrated peak areas of the two bands being approx. 3:1 (± 0.7) in all tracks. No evidence of hybridization to vicilin mRNA species, which have been previously

1988

```

CvcA TCAAG66AACAAATTGAG6AATTGAGAAAGCTTGCAAAATCAAGCTCAAG6AAAAGCTTACCCTCTGAATTTGAACTTTCAACTTGAGAA6CCCAAGCCAGAAATTTCTAATAAGTTT 1534
A.A. S R E Q I E E L R K L A K S S S K K S L P S E F E P F N L R S H K P E Y S N K F
      +-----+
CvcA G6CAAGTTGTTT6AGATTACTCCAGAGAAAAATACCTCAGCTTCAAGATTAGATATACTTGTAGTTGTGTG6AGATTAAACAAGGTATGTACACAACACTAAATATATATAAAACAC 1654
A.A. G K L F E I T P E K K Y P Q L Q D L D I L V S C V E I N K <-----
      +-----+
CvcA ATCATTTTAATTATATTACAGAAATATGTTAATGCTTTTGTCTAAATTTTGA66AGCTCTAATGTTGCCACACTACAATTCAAG66CAATAGTTGTACTATTAGTTAATGAG6AAA 1774
A.A. -----[VS-4]-----> G A L M L P H Y N S R A I V V L L V N E G K
CvcA A66AAACCTTGAACCTTCTG66TTTAAAAATGAGCAACAAGAG66AAGATAGAAAAGAAAGAAACATGAAGTGCAGAGATATGAA6CTAGATTGTCTCC66TGACGTTGTTATCAT 1894
A.A. G N L E L L G L K N E Q Q E R E D R K E R N N E V Q R Y E A R L S P G D V V I I
      +-----+
CvcA TCCAGCAG6TCACCCAGTTGCCATTAGTCTTCATCAAACTCTGAATTTGCTT66ATTGGTATCAATGCCAAGAACATCAGAGAACTTCCTTTTCA6GTATTAGTGAATAGTAATATC 2014
A.A. P A G H P V A I S A S S N L N L L G F G I N A K N N Q R N F L S <-----
CvcA ATAGTTAATAATTTTCGATTAATGAGAAATATTGAATGTTATTTCTAATTTG66GATTGAAAATTTGAAGGATCGGATGACAAATGTGATAAGCCAAATAGAAATCCAGTAAAGG 2134
A.A. -----[VS-5]-----> G S D D N V I S Q I E N P V K
CvcA A6CTCACATTTCTG6ATCTTCTCAAGAG6TAATAGATTATCAAGAAATCAAAACAATCTCACTTTGCAAGTGTGAACTGAAACAAAAG6AG6AAGAAAGCCAAAGAAAAAGGAGTC 2254
A.A. E L T F P G S S Q E V N R L I K N Q K Q S H F A S A E P E Q K E E S Q R K R S
      +-----+
CvcA CTCTGCTTCAGTTCTG6ACAGTTTTTACTGAGTAATCAATATGAAAAAATGAGATGTATGAGCTAAGATCTAGCTAGCTCTTCGTGAGCTAAGAGTAATAATAGGATCTTGTAACT 2374
A.A. P L S S V L D S F Y >.....
CvcA CTTACCTATTGAGCCCCACTTTTCTATACGAATAAATAAATAATTAATAAACTTGTGCTTTTTTTTACTTTAACTACAAGGATAATATTAAATTTGTGTTCTTGTG66GTAAGTCTTAAAA 2494
.....< PolyA+ >...<PolyA+>.....
CvcA AAAGACTATG6ATTCAATGAAG6AATTTTAAAAATTGTTTTTAATAATG6TTATTG6TTGTTATTATTGTAATG6TTTCATAACATGCA6TGCTTACTCTTGTATTAGTTGCTTTAAT 2614
CvcA TTGCTTTAATTTGTTATGTTTTATCTTTTCTTTAAATTAATAAAATGGAAGTGTGTTGTAATTTGTAAGTAAAGACGAG6TTGTGCAATTTCTTTTCTCTCTAGA..... 2723

```

Fig. 2. Sequence of gene *cvcA* ('CvcA'), with the predicted sequence of the convicilin precursor polypeptide ('A.A.')

The predicted site of cleavage of the leader sequence is indicated by a colon (:). The base designated +1 is indicated by a circumflex (^). Other sequence features are as indicated on the Figure. The *N*-terminal sequence determined for convicilin, and the sequences of convicilin tryptic peptides, are indicated by double and single underlinings respectively; vertical lines indicate the termini of the peptides.

identified as approx. 1700 bases in size [33], was obtained, showing that the probe was specific for convicilin mRNA species. The relative intensities of the hybridizing bands from different developmental stages show that the proportion of convicilin mRNA species in total RNA increases as cotyledon expansion proceeds, to a maximum at 16–18 days after flowering, and decreases thereafter. The peak in convicilin mRNA levels agrees with previous observations that convicilin synthesis is maximal during the second half of cotyledon expansion [34].

Hybridization to genomic DNA

Pea genomic DNA from cvs. Feltham First and Dark Skinned Perfection was digested with various restriction enzymes, size-fractionated by agarose-gel electrophoresis

and blotted on to nitrocellulose. The blots were then hybridized with the labelled convicilin specific probe (*Sst*I–*Bgl*II; bases –176 to 462) described above. Results are shown in Fig. 4. The two cultivars gave identical band patterns in all restriction digests made. Digests with *Eco*RI gave two bands, one of approx. 13 kb, corresponding to the *Eco*RI fragment in pJc 4-100, and one of approx. 9.0 kb, corresponding to the *Eco*RI fragment previously identified as hybridizing to the convicilin cDNA species pCD 59 and pCD 75 [5]. Both these bands were present at an indicated level of approx. one copy per haploid genome, as shown by a reconstruction assay where gene copy equivalents of pJc 4-100 were hybridized on the same filters. All other restriction digests gave two or more hybridizing bands, consistent with the restriction

Fig. 3. Expression of convicilin gene *cvcA*

(a) S1 mapping experiment to locate the transcription start in *cvcA*. The protected fragment is run in track S; other tracks are the corresponding region of DNA sequence (the sequence is given in complement, and must be read *down* the sequencing gel). (b) 'Northern' blot, showing hybridization of *Sst*I-*Bgl*II probe (bases -176 to 462) from *cvcA* to total RNA isolated from developing pea cotyledons (line Feltham First) at 8 days after flowering (d.a.f.) (track 1), 10 d.a.f. (track 2), 12 d.a.f. (track 3), 14 d.a.f. (track 4), 16 d.a.f. (track 5), 18 d.a.f. (track 6), 20 d.a.f. (track 7) and 22 d.a.f. (track 8). Under these conditions the cotyledon expansion phase of development lasts from 7-8 d.a.f. to 21-22 d.a.f. [24,32]. A 10 μ g portion of total RNA was loaded per track in the original gel electrophoresis. The molecular-size scale is taken from standard RNA species (ribosomal RNAs) run on the original gel.

map of *cvcA* (see Fig. 1), at intensities consistent with the conclusion that two convicilin genes were present per haploid genome, in agreement with previous reports [6].

DISCUSSION

Coding sequence

The amino acid sequences predicted by *cvcA*, and found for convicilin, confirm the presence of a 'leader' sequence on the precursor polypeptide, as had been previously suggested by translation experiments *in vitro* [35]. The sequence for the mature polypeptide predicted by *cvcA* is then in good agreement with the amino acid composition of convicilin, as shown in Table 1. The presence of one methionine residue in the mature polypeptide is correctly predicted by *cvcA*, and its position (amino acid 388) is consistent with the observed results of CNBr cleavage of convicilin, which generates two fragments of approx. 55000 and 15000 M_r [1].

Despite the evidence that *cvcA* is a convicilin gene and that it is expressed, it differs in its sequence from the convicilin cDNA identified by Domoney & Casey [4], which was used to select the genomic clone containing *cvcA*. The overall homology between the two sequences is 94% over 590 corresponding bases. The main difference between the two sequences is a deletion of 18 nucleotides (six amino acids) in pCD59 relative to *cvcA*, corresponding to a region near the hypothetical $\alpha:\beta$ subunit processing site in vicilin [26]. There are also a number of conservative amino acid substitutions in the remainder of the sequence (not shown). These sequence differences are sufficient to account for the previous observation [5] that pCD 59 hybridized to only one of the two convicilin genes detected by the *cvcA* probe in the present study. The data suggest that pCD 59 represents the second convicilin gene detected by hybridization to genomic DNA, *cvcB*, which is thus shown to be functional. When pCD 59 was hybridized to RNA from developing pea cotyledons [5], only one band was detected

Fig. 4. Southern blot showing hybridization of *SstI*-*BglII* probe (bases -176 to 462) from *cvcA* to restriction digests of genomic DNA from lines Feltham First (F) and Dark Skinned Perfection (D)

A 10 μ g portion of DNA was loaded per track on the original gel electrophoresis. Restriction enzymes used were as follows: A and B, *EcoRI*; D and E, *BglII*; F and G, *BamHI*; H and I *EcoRV*. The blot is calibrated with gene equivalent amounts [33] of digested pJC4-100; the indicated copy numbers per haploid genome are given above tracks C, J, K and L. Tracks A-C are from a different gel to the remainder. The molecular-size scale is from restriction digests of standard DNA species run on the original gels.

on a 'Northern' blot, as opposed to the two detected by the *cvcA* probe, suggesting that *cvcA* and *cvcB* each gives rise to a distinct mRNA species. Further data will be necessary to confirm this conclusion.

Homology with vicilin. A dot-matrix comparison of the polypeptide sequences predicted for convicilin, and for a vicilin 50000- M_r polypeptide is given in Fig. 5. The sequences are strongly homologous over most of their length, with short areas of low homology apparent at regions corresponding to the sequences around the putative $\alpha:\beta$ and $\beta:\gamma$ subunit processing sites in vicilin. These areas have previously been noted as being of low homology when pea vicilin polypeptides are compared with those from different species [28]. The major difference between the two sequences is apparent as a large insertion in the convicilin sequence near its *N*-terminus, corresponding to sequence being inserted between amino acids 3 and 6 of the mature vicilin polypeptide. Homology over the region -3 to +3 is

weak at the amino acid level, but significant at the nucleotide level; outside this region, and the insertion, homology is strong in both directions (see Fig. 5). The convicilin leader sequence is homologous with that in vicilin, but not to leader sequences in other seed proteins (results not shown), showing that the extra sequence in convicilin represents an insertion into a vicilin gene rather than a 5' addition to it. The strong homology of convicilin with vicilin outside the inserted sequence accounts for the overall similarity in properties between the two proteins and their antigenic similarity [1]; it would also account for their ability to form molecules containing polypeptides of both vicilin and convicilin.

The homology in amino acid and corresponding nucleotide sequences between *cvcA* and vicilin genes in pea (results not shown; homology at the nucleotide level between the vicilin cDNA pAD2.1 [29] and corresponding sequence regions in *cvcA* is 79%) shows that the *cvcA* gene should be regarded as belonging to a sub-family of the vicilin gene family; this designation supports both

Table 1. Amino acid composition of convicilin; comparison of predicted and experimental compositions

Amino acid	Residues predicted		Composition (mol/100 mol)	
			Predicted	Found*
D	23	59	10.87	11.64
N	36			
T	13			
S	40	113	7.37	6.39
E	80			
Q	33			
P	25		4.60	5.47
G	27			
A	18			
C	1		0.17	0.13
V	27			
M	1			
I	24		4.42	3.85
L	49			
Y	15			
F	20		3.68	3.30
W	3			
K	43			
H	12		2.21	2.22
R	53			

* From [1].

† ND, not determined.

previous views that convicilin was distinct from [1], or was essentially the same as [3], vicilin.

Nature of the inserted sequence in convicilin. The inserted sequence in convicilin will be considered as amino acids (+)4–124 or nucleotides 121–483. At the amino acid level, the sequence contains a high proportion of charged and hydrophilic residues (from 121 amino acids, there are 38 glutamate residues, 24 arginine residues and 9 lysine residues; only 10 residues are strongly hydrophobic). It is similar in its composition to the C-terminal regions of the α -subunits encoded by both 'major' and 'minor' pea legumin genes ([36,37]; J. A. Gatehouse & D. Bown, unpublished work), but the actual amino acid sequences are not significantly homologous when compared by a dot-matrix homology plot (results not shown). This additional sequence is presumably responsible for the differences in physical properties between vicilin and convicilin, e.g. solubility and binding to hydroxyapatite [1]. The predicted M_r values for the mature convicilin polypeptide, and its N-terminal CNBr fragment, are not in complete agreement with those observed on SDS/polyacrylamide-gel electrophoresis. This discrepancy is a consequence of abnormal migration on electrophoresis, possibly due to the atypical amino acid composition of these polypeptides caused by the 'inserted' sequence.

Fig. 5. Dot-matrix comparison of the amino acid sequences of vicilin (from pAD 2.1 plus *vicB*) and convicilin

Sequences were compared over a span of eight amino acids, with a minimum score of 102 using the correlation matrix given by Staden [15].

At the nucleotide level, the inserted sequence is A + G rich, again like the C-terminal regions of legumin α -subunits; however, overall homology of nucleotide sequence in these regions is not more than marginally significant by dot-matrix comparison. No introns are present in the inserted sequence. There is no evidence of inverted repeats at the ends of the inserted sequence, nor strong evidence for direct repeats in or near the sequence itself (results not shown). The origin of this sequence is therefore unclear; it may represent a sequence inserted by a transposable element or by some other mechanism.

Relationship to vicilin-family genes in other species

The relationships of the coding sequences of vicilins in pea, *Phaseolus vulgaris* (phaseolin) and soya bean (conglycinin) have been extensively analysed, and part of the coding sequence of convicilin has been shown to be homologous with those of phaseolin and conglycinin [38]. Both convicilin and conglycinin have large inserted coding sequences (121 and 174 amino acids respectively) near the N-terminus of the mature protein, relative to the vicilin/phaseolin type. The inserted sequences in convicilin and conglycinin also show similarity at the nucleotide level in that both sequences are A + G-rich. However, the inserted sequences in the two genes are not

significantly homologous at either the amino acid or the nucleotide sequence level. Further, the remaining coding sequences of the two genes, although homologous, are less homologous with each other than convicilin in pea is with pea vicilin, suggesting that the divergence of the pea gene sub-families took place after the separation of pea and soya bean as species. If this is the case, the insertion events were independent of each other. Further analysis of other storage-protein gene sequences (results not shown) suggests that the insertion of hydrophilic, predominantly acidic, amino acid sequence regions is a frequent mechanism of storage protein mutation in legumes.

The flanking sequences

3' Flanking sequence. The 3' flanking sequence of *cvcA* does not show any unusual features when compared with other plant storage-protein genes.

5' Flanking sequence. Features of potential interest in the 5' flanking sequence of *cvcA* were shown by dot-matrix sequence comparisons between this gene and other plant storage-protein genes. Comparisons of the 5' flanking sequence of *cvcA* with those of conglycinin and phaseolin genes show three areas of sequence con-

'Vicilin-box' region

```

                : v ( 106 ) :
Pvu phas b    CC : GCCACCTCAATTC - TCACTTCAACACACGTC AACCTGCAT : AT
                : v ( 88 ) :
Gma cgly a'   CC : GCCACCTCATTTT TGT TTTATTTCAACACCCGTC AA ACTGCAT : CC
                : v ( 99 ) :
Psa cvcA      TT : GCCACCTCTATTT TGT TTTCACTTCAACACTCGTCAAGTTACAT : GA
                : ***** ** * ***** * *** :
                : ^ :
                ^ (distance to 'TATA' box)

```

Upstream region 1

```

                : v ( 180 ) :
Pvu phas b    GGC : TCACCCATCTCAACCC : ACAC
                : v ( 153 ) :
Gma cgly a'   CAT : TCAC - CAACTCAACCC : ATCA
                : v ( 152 ) :
Psa cvcA      TAA : TCAA - CAACTCAACCC : GCGA
                : *** ** ***** :
                : ^ :
                ^ (distance to 'TATA' box)

```

Upstream region 2

```

                : v ( 257 ) :
Pvu phas b    GGC : TGATCAAGATCGCCGCTCCA : TGTATG
                : v ( 251 ) :
Gma cgly a'   AGC : TGATCAGGATCGCCGCTCAA : GAAAAA
                : v ( 255 ) :
Psa cvcA      TCA : TGGTCATGATCGCCGCTCCA : TGTAAA
                : ** *** ***** ** * :
                : ^ :
                ^ (distance to 'TATA' box)

```

Fig. 6. Putative enhancer sequences in the 5' flanking regions of *cvcA*

The three corresponding regions of high sequence homology between pea convicilin (*Psa cvcA*), *Phaseolus vulgaris* phaseolin b (*Pvu phas b*) and soya-bean conglycinin a' (*Gma cgly a'*) gene 5' flanking sequences are given. Bases the same in all three sequences are indicated by an asterisk. Homologous regions around the transcription start and the 'TATA' box are not shown.

servation besides the 'TATA' box promoter element (considered previously); the conserved regions are shown in Fig. 6. There is also a conserved region around the transcription start, which has an obvious functional role, and a possible further conserved region of approx. 15 bases, at 30–50 bases 5' to the 'TATA' box. This latter region is not as well conserved or defined as other regions, but does include the putative CCAAT sequences of phaseolin and conglycinin [39].

The 'vicilin box' region [39] in all three genes is in a similar position (approx. 100 bases 5' to the 'TATA' box), and is strongly homologous; it can be divided into two regions, separated by 11–12 bases of T-rich sequence. The 5' region is a highly conserved C-rich sequence (GCCACCTC), whereas the 3' region is more typical of the 5' flanking sequence as a whole (TTCAACACNCGTCAANNTG/ACAT). It has been suggested that this region, present also in pea vicilin genes, is involved in determining tissue-specificity of expression of the gene family [39]. The other two conserved regions are approx. 150–200 bases and 250 bases 5' to the 'TATA' box; like the 'vicilin box', both seem to have a highly conserved C-rich core sequence (CTCAACCC and GATCGCCGC respectively) and are associated with less highly conserved sequence more typical of the 5' flanking sequence as a whole. The hypothesis that such C-rich sequences are acting as 'enhancers' of gene expression may be advanced, and is supported by the observation that the 'vicilin-box' C-rich sequence is present in the pea legumin gene *legA* also, and has been previously observed to be homologous with a viral enhancer sequence [39,40]. However, functional assays such as those carried out with the conglycinin gene in transgenic petunia plants [41] are needed to test this conclusion.

We thank Dr. H. Hirano, National Institute of Agro-biological Resources, Tsukuba, Japan, for carrying out the automated protein sequencing, John Gilroy for performing manual protein sequencing, and Paul Preston for skilled technical assistance in DNA sequencing. We also thank Professor D. Boulter for providing departmental facilities. Financial support from the Agriculture and Food Research Council and the Science and Engineering Research Council is gratefully acknowledged.

REFERENCES

1. Croy, R. R. D., Gatehouse, J. A., Tyler, M. & Boulter, D. (1980) *Biochem. J.* **191**, 509–516
2. Croy, R. R. D. & Gatehouse, J. A. (1985) in *Plant Genetic Engineering* (Dodds, J. H., ed.), pp. 143–268, Cambridge University Press, Cambridge
3. Thomson, J. A. & Schroeder, H. E. (1978) *Aust. J. Plant Physiol.* **5**, 281–294
4. Domoney, C. & Casey, R. (1983) *Planta* **159**, 446–453
5. Casey, R., Domoney, C. & Stanley, J. (1984) *Biochem. J.* **224**, 661–666
6. Domoney, C. & Casey, R. (1985) *Nucleic Acids Res.* **13**, 687–699
7. Matta, N. K. & Gatehouse, J. A. (1982) *Heredity* **48**, 383–392
8. Mahmoud, S. H. & Gatehouse, J. A. (1984) *Heredity* **53**, 185–191
9. Ellis, T. H. N., Domoney, C., Castleton, J., Cleary, W. & Davies, D. R. (1986) *Mol. Gen. Genet.* **205**, 164–169
10. Maniatis, T., Fritsch, E. F. & Sambrook, J. (1982) *Molecular Cloning — A Laboratory Manual*, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY
11. Vieira, J. & Messing, J. (1982) *Gene* **19**, 259–268
12. Messing, J. (1983) *Methods Enzymol.* **101**, 20–78
13. Sanger, F., Nicklen, S. & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. U.S.A.* **74**, 5463–5467
14. Biggin, M. D., Gibson, T. J. & Hong, G. F. (1983) *Proc. Natl. Acad. Sci. U.S.A.* **80**, 3963–3965
15. Staden, R. (1982) *Nucleic Acids Res.* **10**, 295–306
16. Hopp, T. R. & Wood, K. R. (1981) *Proc. Natl. Acad. Sci. U.S.A.* **78**, 3824–3828
17. Kuhn, S., Anitz, H. J. & Starlinger, P. (1979) *Mol. Gen. Genet.* **167**, 235–241
18. Rigby, P. W. J., Dieckmann, M., Rhodes, D. & Berg, P. (1977) *J. Mol. Biol.* **113**, 237–251
19. Ellis, T. H. N., Davies, D. R., Castleton, J. A. & Bedford, I. D. (1984) *Chromosoma* **91**, 74–81
20. McMaster, G. K. & Carmichael, G. G. (1977) *Proc. Natl. Acad. Sci. U.S.A.* **74**, 4835–4838
21. Chirgwin, J. M., Przybyla, A. E., Macdonald, R. J. & Rutter, W. J. (1979) *Biochemistry* **18**, 5294–5299
22. Thomas, P. S. (1980) *Proc. Natl. Acad. Sci. U.S.A.* **77**, 5202–5205
23. Favaloro, I., Treisman, R. & Kamen, R. (1980) *Methods Enzymol.* **65**, 718–749
24. Gatehouse, J. A., Evans, I. M., Bown, D., Croy, R. R. D. & Boulter, D. (1982) *Biochem. J.* **208**, 119–127
25. Maxam, A. M. & Gilbert, W. (1980) *Methods Enzymol.* **65**, 499–560
26. Gatehouse, J. A., Lycett, G. W., Croy, R. R. D. & Boulter, D. (1982) *Biochem. J.* **207**, 629–632
27. Slightom, J. L., Sun, S. M. & Hall, T. C. (1983) *Proc. Natl. Acad. Sci. U.S.A.* **80**, 1897–1901
28. Lycett, G. W., Delauney, A. J., Gatehouse, J. A., Gilroy, J., Croy, R. R. D. & Boulter, D. (1983) *Nucleic Acids Res.* **11**, 2367–2380
29. Delauney, A. J. (1984) Ph.D. Thesis, University of Durham
30. Lycett, G. W., Delauney, A. J. & Croy, R. R. D. (1983) *FEBS Lett.* **153**, 43–46
31. 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
32. Von Heijne, G. (1985) *J. Mol. Biol.* **184**, 99–105
33. Croy, R. R. D., Lycett, G. W., Gatehouse, J. A., Yarwood, J. N. & Boulter, D. (1982) *Nature (London)* **285**, 76–79
34. Tyler, M. (1981) Ph.D. Thesis, University of Durham
35. Higgins, T. J. V. & Spencer, D. (1981) *Plant Physiol.* **67**, 205–211
36. Lycett, G. W., Croy, R. R. D., Shirsat, A. & Boulter, D. (1984) *Nucleic Acids Res.* **12**, 4493–4506
37. Gatehouse, J. A., Bown, D., Gilroy, J., Levasseur, M., Castleton, J. & Ellis, T. H. N. (1988) *Biochem. J.* **250**, 15–24
38. Doyle, J. J., Schuler, M. A., Godette, W. D., Zenger, V., Beachy, R. N. & Slightom, J. L. (1986) *J. Biol. Chem.* **261**, 9228–9238
39. Gatehouse, J. A., Evans, I. M., Croy, R. R. D. & Boulter, D. (1986) *Philos. Trans. R. Soc. London B* **314**, 367–384
40. Lycett, G. W., Croy, R. R. D., Shirsat, A. H., Richards, D. M. & Boulter, D. (1985) *Nucleic Acids Res.* **13**, 6733–6743
41. Chen, Z.-L., Schuler, M. A. & Beachey, R. N. (1986) *Proc. Natl. Acad. Sci. U.S.A.* **83**, 8560–8564