Characterisation of genes expressed in various tissues of PEA (Pisum sativum L.); correlation of genotype and phenotype

David Philip Bown

CHARACTERISATION OF GENES EXPRESSED IN VARIOUS TISSUES OF PEA (*PISUM SATIVUM* L.); CORRELATION OF GENOTYPE AND PHENOTYPE

A thesis submitted by David Philip Bown, B. Sc. (Newcastle) in accordance with the requirements of the University of Durham for the degree of Doctor of Philosophy.

Department of Biological Sciences

June 1992

ABSTRACT

Genes encoding representatives of two subfamilies from the vicilin storage protein gene family in pea (*Pisum sativum* L.) have been sequenced and characterised. One, encoding convicilin, shows that this protein differs from vicilin by the insertion of a hydrophilic region near the N-terminus. The transcription start point has been determined and the pattern of expression in developing seeds elucidated. By the expression of this gene in tobacco, the specific polypeptide product of the gene was identified as a minor component of convicilin, with a lower Mr than the major species. The other gene subfamily investigated was that encoding the vicilin 47,000 Mr polypeptide. A gene and a cDNA were sequenced, and the gene found to diverge from the cDNA in the 3' region of the coding sequence. No product from this divergent gene could be identified.

A member of the legumin gene family (*legK*) was sequenced and found to be inactive due to a mutation of the start codon. The region of DNA encoding the start codon of this gene was amplified by polymerase chain reaction from a pea line in which the gene was known to be active. The sequence of this revealed the presence of a normal start codon. Two-dimensional protein gels were run with seed extracts from these two lines, and the product of (*legK*) demonstrated by its occurrence in the line with the functional gene.

A method for the extraction and purification of the major pea root protein was established. The protein was shown to have a Mr of 16,000 and not to be susceptible to cleavage by cyanogen bromide. Partial amino acid sequence data was obtained from the purified protein.

A differential screen of cDNA from purple and green podded varieties of pea was conducted, and differentially expressed cDNAs isolated. The nature of the expression of these cDNAs was studied in the two lines and the cause of instability in the purple podded phenotype investigated. A genomic library was constructed from the purple podded line. Two genes were selected by the differentially expressed cDNAs and their DNA sequences determined. A gene encoding a pectinesterase-like sequence, and the pod expressed cDNA used to select it, were found to be two members of a small multigene family in pea. The second gene selected proved to encode a protein containing two distinct domains; the N-terminal region being of a repetitive proline-rich nature and the C-terminal region being hydrophobic and cysteine rich. This gene was present as a single copy in the pea genome and its expression appeared to be linked to pigmentation.

CONTENTS

FIGURES

MEMORANDUM

Parts of this work have been included in the following publications (see appendix):

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

Thompson,AJ., Bown,D., Yaish,S. and Gatehouse,JA. (1991). Differential expression of seed storage protein genes in the pea *legJ* subfamily; sequence of gene *legK*. Biochem. Physiol. Pflanzen *187*, 1-12.

ABBREVIATIONS

aa	amino acid
amp	ampicilin
ATP	adenosine-5'-triphosphate
b(p)	base (pair)
BSA	bovine serum albumin
cDNA	copy DNA
CHS	chalcone synthase
CoA	coenzyme A
cpm	counts per minute
C-terminal	carboxy-terminal
cv	cultivar
DABITC	4-N,N-dimethylaminobenzene-4'-isothiocyanate
daf	days after flowering
dATP	deoxyadenosine-5'-triphosphate
dCTP	deoxycytidine-5'-triphosphate
dGTP	deoxyguanosine-5'-triphosphate
DNA	deoxyribonucleic acid
DNase	deoxyribonuclease
DSP	Dark Skinned Perfection
DTT	dithiothreitol
dTTP	deoxythymidine-5'-triphosphate
EDTA	ethylene diamine tetra-acetic acid
ELISA	enzyme linked immunosorbant assay
FF	Feltham First
GP	Green podded mutants from PP line
HPLC	high performance liquid chromatography
IAA	iso amyl alcohol
IPTG	isopropyl-β-thiogalactopyranoside

k	$x10^3$
kD	kiloDalton
2-ME/β-ME	2/β -mercaptoethanol
MOPS	3-(N-morpholino) propane-sulphonic acid
Mr	relative molecular weight
mRNA	messenger RNA
N-terminal	amino-terminal
NBRF	National Biomedical Research Foundation
OD_{260}	Optical density at 260nm
oligo d(T)	oligodeoxythymidylic acid
ORF	open reading frame
PAGE	polyacrylamide gel electrophoresis
PBS	phosphate buffered saline
PCR	polymerase chain reaction
pfu	plaque forming units
PIPES	piperazine-N,N'-bis(2-ethane-sulphonic acid)
PMSF	phenylmethylsulphonyl flouride
poly(A)	polyadenylic acid
PP	Purple Podded
ppt	precipitate
RNA	ribonucleic acid
RNase	ribonuclease
rpm	revolutions per minute
SDS	sodium dodecyl sulphate
SSC	saline sodium citrate
TBS	Tris buffered saline
TCA	trichloroacetic acid
TE	Tris-EDTA
Tris	tris(hydroxymethyl)methylamine

tRNA	transfer RNA
UV	ultra violet light
X-gal	5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside

Single letter abbreviations for amino acids and bases are as specified in; Biochem. J. (1984) *219*, 345-373 and Biochem J. (1985) *229*, 281-286. respectively.

ACKNOWLEDGEMENTS

I would like to thank the following for donating materials used in this project: Dr.THN.Ellis for the genomic clone and subclones containing *cvcA* and *vicJ*; Dr.GA.Edwards for pDUD126A vector; Mr.LN. Gatehouse for pLG1.63 cDNA; Dr.CJ.Lamb for *Phaseolus* Chalcone synthase and Chalcone isomerase cDNAs; The Nordic Gene Bank for pea seed material; Dr.C.O'Rielly for maize *A1* genomic DNA, Dr.GA.Tucker for anti-tomato pectinesterase antiserum; Dr.SA.Yaish for the genomic subclone containing *legK*.

I am very grateful to Mr.Niel Appleby for technical assistance, Ms.Julia Bryden for DNA sequencing and PCR amplification of the *legK* gene from Birte, Mr.John Gilroy for oligonucleotide synthesis and protein sequencing, and Mr.Paul (Doc) Preston for DNA sequencing.

I would like to thank the following members of the department for allowing me to use their unpublished results; Ms.Julia Bryden, Dr.Ron Croy, Dr.John Gatehouse, Mr.John Gilroy, Mr.Dave Jobes, and Ms.Jackie Spence, Mr.John Davies and Ms.Leslie Edwards for allowing me to use their pectinesterase *in situ* hybridisation results.

I am indebted to Dr.Ron Croy, Dr.Marta Evans and Mr.Russell Swinhoe for helpful discussion and advice on laboratory techniques. Thanks are also due to Prof.D.Boulter for his encouragement and support throughout this project.

I am especially grateful to Dr.John Gatehouse, not only for his supervision of this project, but for his guidance and encouragement, without which I would not have been in a position to have embarked upon this work. Finally, I would like to thank Ms.Heather Edmonds for sustaining me, and my garden, during the course of writing this thesis and for proof reading the manuscript.

CHAPTER ONE: INTRODUCTION

General introduction and aims of the project

All the cells within a given plant originate from a single embryo cell and throughout subsequent division and growth all retain the same basic genetic information. However, the plant is composed of numerous differentiated cell types in the various separate organs, and within these organs, in specific tissue types. The differences between these cell types result from differential expression of the genetic material within them, ie. in the qualitative and quantitative differences in the mRNA populations in the cells (this ignores cytoplasmic proteins passed from mother to daughter cells and organelles - chloroplasts and mitochondria). mRNA produced within plant cells can be divided into three classes (Goldberg, 1986); mRNA sequences shared between all cells, encoding the proteins required by all cells; sequences specific to a particular organ and sequences specific to two or more organs. An investigation of genes encoding the latter types of mRNA formed the basis of this project. Three different organs from pea were chosen for study in this work, one, the seed, has been the subject of a great deal of previous work, whereas the other two, the pod and the root have been comparatively little studied by the methods of molecular biology. The project aimed to isolate and characterise tissue-specific genes from these organs, and to investigate the connection between gene and phenotype.

Work on the pea seed has been mainly concentrated on genes encoding the storage proteins. These are expressed in a tissue and developmental specific manner. By assembling a body of nucleotide sequence data from the flanking regions of these genes, and similar genes from related genera, it should be possible to identify sequences which have been conserved through their evolution. The conservation of such sequences

implies a functional significance for them which can then be tested in transgenic plants. The transfer of genetic material into transgenic plants has allowed the assumption that *cis*-acting elements within the 5' flanking regions of genes convey the information required for tissue-specific expression to be confirmed. Analysis of these regions, combined with functional testing through plant transformation with deletion mutants and reporter gene fusions, should yield precise information on the nature of these elements. Conservation of these sequences may be expected on a family-wide basis as far as tissue specificity goes, but both the major storage proteins exhibit differential developmental expression of their subfamilies. This suggests the possibility of sequences conserved within subfamilies (but not within the family as a whole) and even between subfamilies of two different families whose encoded proteins share the same developmental pattern.

The complex phenotype of the pea seed storage proteins has been extensively studied (see below). Post-translational processing plays an important part in determining the phenotype of storage protein polypeptides. Elucidation of the relationships between storage protein polypeptides and their precursors, and their respective coding genes should serve as a model for other proteins. This aspect is of particular relevance when "foreign" genes are expressed in transgenic plants under heterologous promoters, as correct processing of translation products may be required for optimal expression.

The purple podded phenotype presented an example of series of genes (encoding the anthocyanin biosynthesis pathway) at least some of which were being expressed in a tissue specific manner, and which might be isolated by differential screening. Isolation of a structural or regulatory gene from this pathway would be interesting in itself, but the presence of instability in the production of pigmentation in this

tissue added an extra incentive to investigate this phenotype. Although little work has been conducted on either pod or root tissue, specific expression in these tissues may well be required of genes introduced into plants to confer, for example, pest or fungal resistance.

Pea seed storage proteins and their coding genes

Legume and cereal seed proteins were classified by Osborne in 1924 on the basis of their solubility, and although the original criteria used have been modified slightly (Croy & Gatehouse, 1985), this classification is still used. The four classes are: 1) Albumins - proteins soluble in water or low ionic strength buffer; 2) Globulins - soluble in dilute salt solution at around neutral pH; 3) Prolamins - soluble in alcohol; 4) Glutelins - soluble in Alkali or strongly denaturing solvent.

Seed storage proteins may be defined as those proteins which break down rapidly after seed germination to provide nutrient for the developing seedling (Derbyshire *et al.*,1976). In pea, as in other legumes, the major seed storage proteins occur in the globulin fraction (Gatehouse *et al.*,1984). This globulin fraction may be divided into two groups with sedimentation coefficients of 11S - the legumins, and 7S - vicilins. The ratio of vicilin to legumin varies between legume species, between varieties within species (Derbyshire, *et al.*,1976) and also through seed development (Gatehouse *et al.*,1982b).

Pea seed development may be divided into two phases (fig 1). During the first phase, up to around day eight after flowering, cell division occurs but little storage material is accumulated. In the second phase (8-21 daf) cell division is superseded by cell expansion and deposition of storage material (starch and protein) takes place. These two phases are followed by developmental arrest and seed desication.

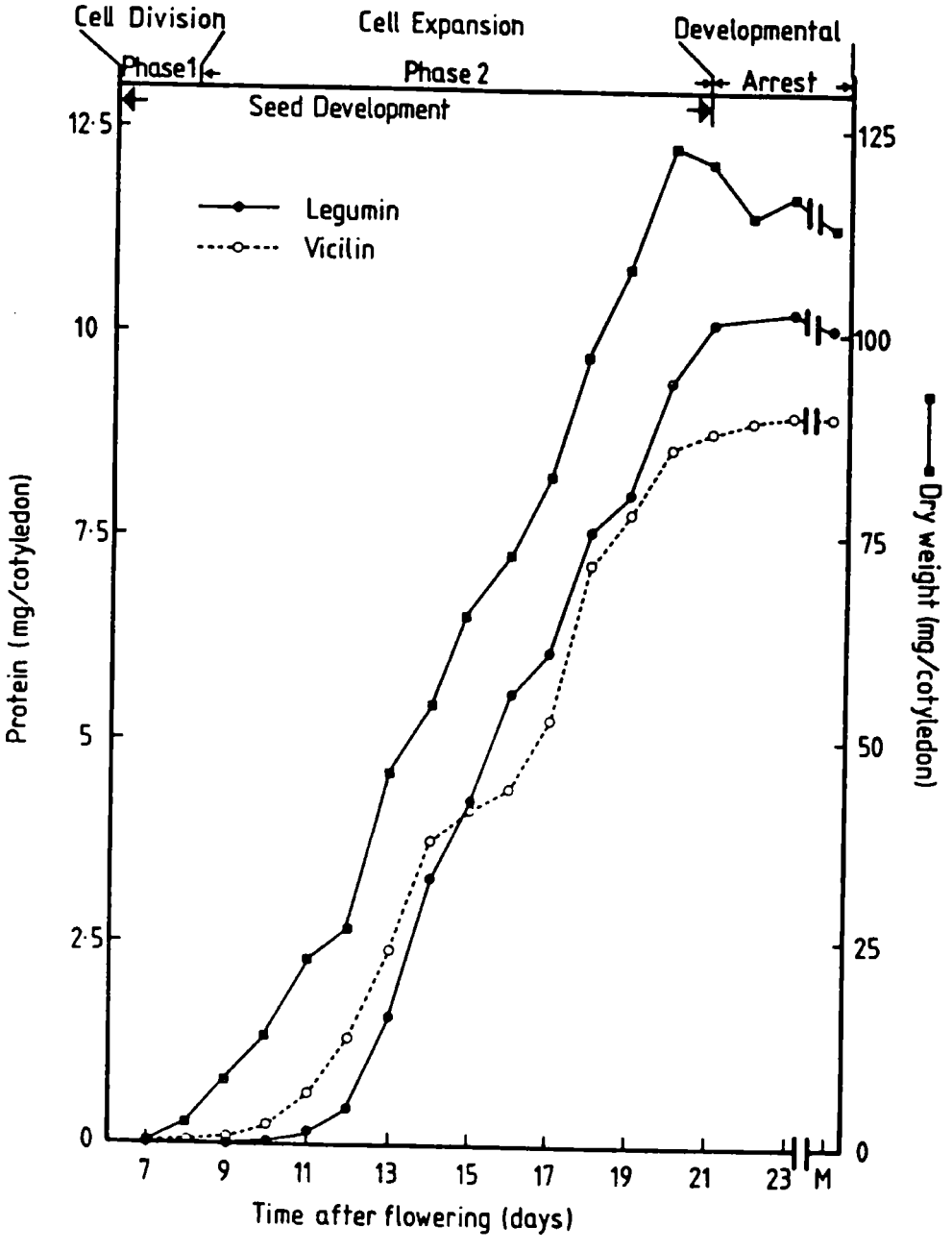


Fig. 1 Vicilin, legumin and dry weight levels in the developing Feltham First cotyledon. From Gatehouse *et al.*,1982b.

Pea Legumin

1) The protein

Pea legumin is present in its native state as oligomers of Mr 330-410,000 (Derbyshire *et al.*,1976, Casey, 1979, Croy *et al.*,1979), which are composed of disulphide bonded subunit pairs with a combined Mr of approximately 60,000 and are arranged as heterogeneous hexamers (Derbyshire *et al.*,1976, Matta *et al.*,1981). The legumin subunit pair consists of an α-subunit of Mr around 40,000 with a pI in the range 4.8-6.1 (also referred to as the "acidic" subunit) and a β-subunit of Mr around 20,000 and a pI in the range 6.2-8.0 (the "basic" subunit) (Matta *et al.*,1981). The heterogeneity of these subunits leads to a complex pattern when they are resolved by two dimensional electrophoresis (fig. 23, Casey, 1979, Krishna *et al.*,1979). However, it was shown that each type of α-subunit was paired with a specific β-subunit and that assembly of these pairs into oligomers was not a random process (Matta *et al.*,1981).

The legumin subunit pairs can be divided into two categories on the basis of their α-subunit Mr and pI (Casey, 1979). These categories comprise; the "major" pairs, a homogeneous, abundant group labelled *legA-C* on fig. 23 (corresponding to Casey's α^M subunits with their respective β-subunits, or the L4 pair of Matta *et al.*,1981), and the remaining subunit pairs, the "minor" subunits (or α^m with their β-subunits), more dispersed in terms of Mr and pI. The minor subunit pairs were further classified in terms of their α-subunit Mr into "big", Mr greater than the α^M-subunit (corresponding to L1-3 of Matta and *legJ,K & X* on fig. 23), and "small", Mr lower than the α^M-subunit and of around 25,000 (corresponding to L5 of Matta and *legS* on fig. 23) (Matta *et al.*,1981).

2) Genetics

On the basis of segregation analysis on the various subunit pairs, legumin genes were mapped to three loci; *Lg-1* the major subunits which segregated as a single locus, *Lg-2* the big minor subunits, and *Lg-3* the small minor subunits. (Thomson & Schroeder, 1978, Casey, 1979, Matta & Gatehouse, 1982). When cDNAs encoding these legumin subfamilies were isolated these results were confirmed (Domoney *et al.*,1986) and genomic clones were later isolated containing two pairs of genes within 2kbp (*legA* & *legD*, Bown *et al.*,1985) and 7kbp (*legJ* & *legK*, Gatehouse *et al.*, 1988) of each other.

3) The genes

As amino acid and nucleotide sequence data became available it became possible to confirm the initial work carried out on the protein and to correlate the legumin phenotype to its genotype. Protein sequence data was obtained from the major subunit pair (Casey *et al.*,1981a & 1981b, Croy *et al.*,1982) and five corresponding genes were subsequently isolated and sequenced (Lycett *et al.*, 1984a & 1985, Mahmoud, 1985 *legA-C*; Bown,*et al.*,1985, *legD*; Rerie *et al.*,1990, Yaish,1990, *legE*), although one of these, *legD*, was found to be a pseudogene.

The proteins encoded by these genes were found to consist of a leader sequence followed by the α-subunit and in the C-terminal region the β-subunit. Cysteine residues were present in both subunits, which accounted for the disulphide bonding, the remaining aa composition being unremarkable except for the presence of a highly acidic region at the C-terminal end of the α-subunit. The genes within this subfamily proved to be highly homologous to each other, the only significant variability being in the length of the acidic region mentioned above.

Protein sequence from the L2 subunit pair of the big minor legumin subunits was found to correspond to that encoded by two genes *legJ* and

legK (Gatehouse *et al.*,1988), whereas no nucleotide sequence encoding the protein sequence from the L1 subunits of big minor legumin (March *et al.*,1988, *LegX* on fig. 23) has been isolated. cDNAs have been isolated which encode the three subunits of small minor legumin (*legS* on fig. 23, Domoney, *et al.*,1986b, Gatehouse,JA., & Gilroy,JS., unpublished results).

4) Legumin expression and biosynthesis

The tissue and developmental control of legumin gene expression has been shown to be regulated primarily at the transcriptional level (Evans *et al.*,1984, Thompson *et al.*,1989). Legumin accumulates in the seed during the latter period of cotyledon expansion (fig. 1, Gatehouse *et al.*,1982b) and this accumulation can be seen to be the result of the onset of legumin mRNA production (Gatehouse *et al.*,1986). However, there is variation between the timing of expression of the different legumin genes. Whilst the expression of the *legA* gene family and *legL* increases to a maximum between 16-20 daf and then decreases during cotyledon desiccation, *legS* mRNA levels peak at an earlier stage during development (16daf) and *legJ* later (22daf) (Domoney & Casey,1987, Thompson *et al.*,1989 & 1991).

Legumins are translated as precursor proteins of Mr around 60-65,000 and 80,000, these precursor proteins are not reducible into smaller subunits (Croy *et al.*,1980b, Chrispeels *et al.*,1982, Domoney & Casey,1984). It appears that these precursors, synthesised in the endoplasmic reticulum, are assembled into oligomers and then transported to the protein bodies where proteolytic cleavage occurs (Chrispeels *et al.*,1982, Gatehouse *et al.*,1984).

Pea Vicilin

1) The protein

The pea 7S storage proteins, the vicilins, have a Mr of 150-190,000. When denatured with SDS, vicilin polypeptides with Mr 50,000, 35,000, 33,000, 30,000 19,000 16,000, 13,500 and 12,500 are observed (Thomson *et al.*,1980, Gatehouse *et al.*,1981). Unlike legumin, the vicilin subunit Mrs are unaffected by reducing agents, and some were shown to contain carbohydrate (Davey & Dudman, 1979, Gatehouse *et al.*,1980). Another distinguishing feature of vicilin from legumin is that it lacks sulphur amino acids (Croy *et al.*,1980a).

A separable protein which cross-reacted antigenically with vicilin was characterised. This protein also appeared in crude vicilin preparations, and so was termed convicilin. It has a native Mr of around 280,000 and a subunit Mr of 71,000 suggesting it forms a simple tetramer (Croy *et al.*,1980a). Unlike vicilin, the aa composition data for convicilin predicted that it contained one cysteine and one methionine residue per subunit (Croy *et al.*,1980a).

Both vicilin and convicilin show considerable variability between pea lines and as many as 30 vicilin subunit species could be separated on two dimensional isoelectric focusing/SDS· polyacrylamide gels (Gatehouse *et al.*,1984). Because of this complexity, further resolution of the vicilin subunit relationships had to await aa and nucleotide sequence data.

2) Vicilin subunit relationships

In vitro translation and pulse chase labelling had revealed that anti-vicilin immunoprecipitable protein precursors were synthesised primarily as 70,000, 50,000 or 47,000 Mr molecules of which a proportion of the 50,000 and all the 47,000 Mr species were subsequently cleaved to yeild the lower Mr species. From these results it was proposed that the

vicilin molecules, like legumin, were synthesised as precursors, assembled in to oligomers (in the case of vicilin, trimers), and exported into protein bodies with the simultaneous removal of the signal peptide. In the protein bodies the vicilin is then subject to proteolytic cleavage at a fairly slow rate (Gatehouse *et al.*,1981, Chrispeels *et al.*,1982).

When cDNA sequence and amino acid subunit sequence became available the conclusions previously drawn were confirmed. The subunits were shown to be the result of differential proteolytic cleavage of a precursor "model molecule" which contained two potential cleavage sites with three polypeptide subunits α:β:γ of Mr 19,000, 13,500 and 12,000. Different cDNAs were found to encode different sequences at the potential cleavage sites and it appears that this sequence determines whether cleavage occurs. The observed subunit Mr can then be accounted for by cleavage, partial cleavage, or non-cleavage of the precursor vicilin molecule: 50,000 - no cleavage, 35,000 & 33,000 - partial cleavage α+β, 30,000 - partial cleavage β+γ, 19,000 13,500 and 12,000 complete cleavage (Gatehouse *et al.*,1982a, 1983 & 1984, Spencer *et al.*,1983). The 16,000 Mr subunit was found to be the result of glycosylation (previously observed in the 16k subunit) of the 12,000 Mr γ-subunit with only some cDNA species encoding the N-glycosylation sequence, N-A-S (Gatehouse *et al.*,1984).

3) Vicilin genes

Three classes of vicilin cDNAs were originally defined on the basis of their cross-hybridisation and the products of translation of hybrid-selected mRNAs (Croy *et al.*,1982, Domoney & Casey, 1983, Ellis *et al.*,1986). These three classes encoded the 47,000 and 50,000 Mr vicilin precursors and the convicilin 70,000 Mr precursor. Hybridisation of these cDNAs to genomic DNA revealed the presence of 5-7 genes encoding

the 47k vicilin, 4-6 encoding the 50k vicilin and 1-3 encoding convicilin (Domoney & Casey, 1985).

Recently, a further class of vicilin cDNA has been isolated and sequenced (Domoney & Casey, 1990). This cDNA hybrid-selects an mRNA encoding a 68,000 Mr precursor which is precipitable by anti-vicilin antibodies. The aa sequence predicted by this cDNA is only 33% identical to that predicted by the closest of the species previously isolated, and differs from the other vicilin classes by being methionine rich. It appears that this cDNA encodes a minor, methionine containing, vicilin species, possibly that previously identified by Chrispeels *et al.* (1982). Although of a minor nature as judged by protein composition, genomic hybridisation suggests that this species is a member of a three gene subfamily with yet another, distinct subfamily related to it (Domoney & Casey, 1990).

Despite the large number of vicilin genes shown to exist, few have been successfully isolated and sequenced. In this laboratory, a portion of a 50k Mr vicilin gene was isolated and sequenced (Sawyer, 1986) together with two pseudogenes (Boulter, *et al.*,1990, Gatehouse JA., personal communication). One full length vicilin gene encoding the 50k vicilin has been isolated and shown to be functional in transformed tobacco (Higgins *et al.*,1988) and two convicilin genes have been sequenced and expressed in tobacco (Bown *et al.*,1988, this work, Newbigin *et al.*,1990).

Early work on the genetics of vicilin coding genes defined two loci for vicilin and one for convicilin (Thomson & Schroeder, 1978, Matta & Gatehouse,1982). More recently the number of vicilin loci has been increased to five with each locus containing multiple genes with greater homology within, rather than between loci (Ellis *et al.*,1986). The convicilin coding genes remain linked in one locus (Ellis *et al.*,1986).

4) Vicilin gene expression

There is differential developmental expression of the vicilin gene subfamilies. Vicilin 50k and 47k encoding mRNA and protein accumulates earlier during cotyledon development than legumin and the other vicilin subfamilies. 50k and 47k vicilin encoding mRNAs reach peak levels at 14-16 daf and decrease markedly to 22 daf. Convicilin and the methionine-rich vicilin accumulate in the final stages of cotyledon developement (Gatehouse *et al.*,1984, Domoney & Casey, 1990).

Genes expressed in pea pods - The purple podded phenotype

The purple podded phenotype in peas has been studied since the beginning of this century (see Blixt,1962 & 1976) and has been shown to be the result of two dominant, polymeric genes, *Pur* and *Pu* (Lamprecht, 1953). Colour production in the pod, as in the other tissues, is also dependent on dominance at the *a* locus which governs overall flavonoid production in the plant (Statham *et al.*,1972). This locus has recently been shown to regulate the expression of chalcone synthase (Harker *et al.*,1990), the initial enzyme of the flavonoid specific biosynthesis pathway. Purple pods have been analysed and shown to contain glycosides of the anthocyanidins delphinidin and cyanidin (Statham *et al.*,1972). Anthocyanins are phenolic compounds within the overall group known as flavonoids.

The anthocyanin biosynthesis pathway

The anthocyanin biosynthesis pathway has been well characterised (Fig. 2, Ebel & Hahlbrock, 1982, Heller, 1986, Heller & Forkmann, 1988). Several genes have been isolated encoding enzymes from this pathway, principally from *Antirrhinum*, maize, and *Petunia* and their associated phenotypes have been described (see Coen *et al.*,1988 & Dooner *et*

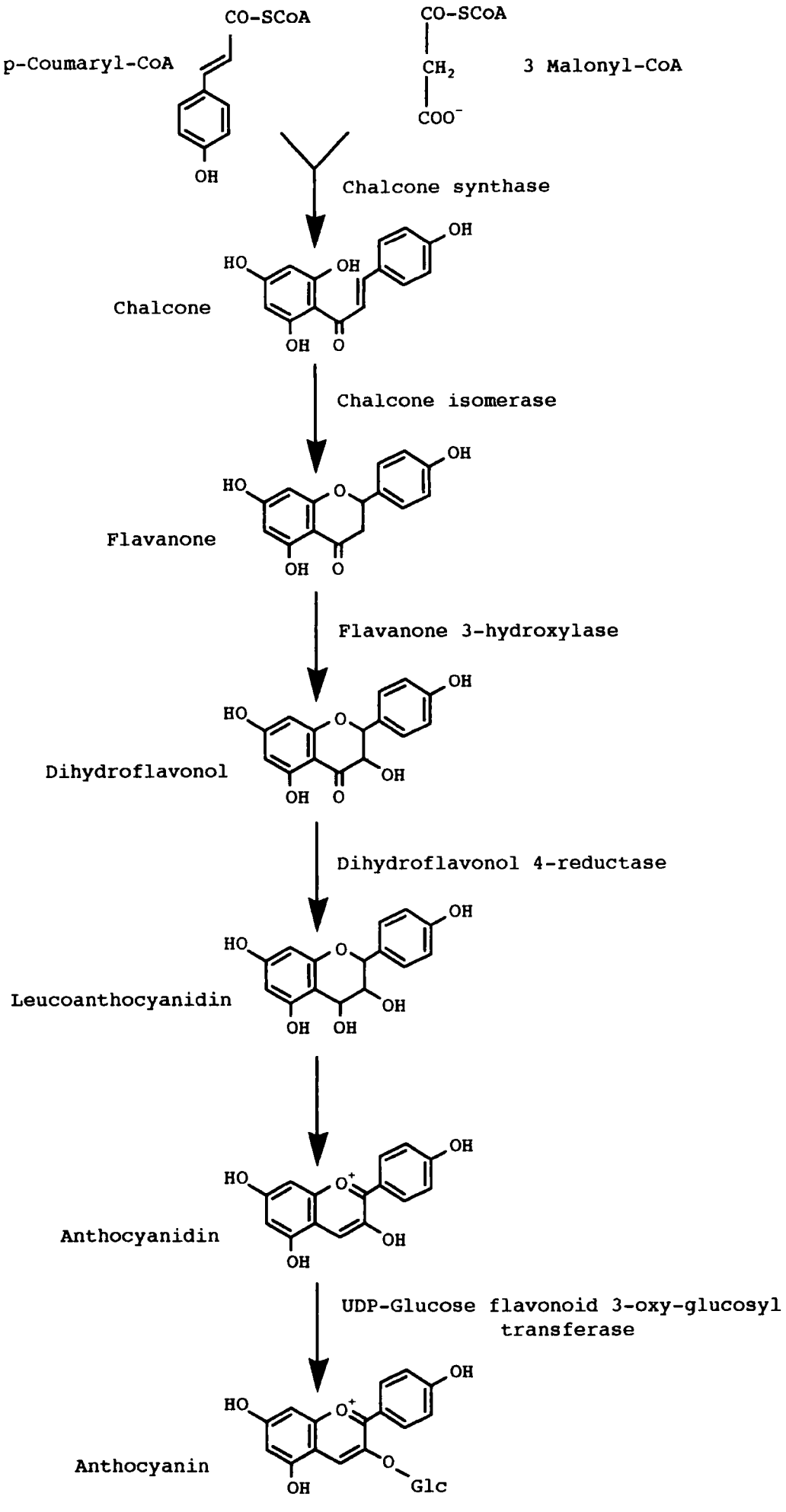


Fig. 2 An outline of the anthocyanin biosynthesis pathway

al.,1991 for reviews).

As stated above, the initial flavonoid specific step is catalysed by chalcone synthase and involves the condensation of coumaryl CoA (derived fron phenylalanine) and malonyl CoA molecules to form a chalcone. The chalcone is then rapidly isomerised in the presence of chalcone isomerase to a colourless flavanone. Subsequent modification of the flavanone by hydroxylation and reduction results in the still colourless leucoanthocyanidin. The next step in the pathway is not well characterised but results in a coloured anthocyanidin. Finally, the anthocyanidin is glycosylated to anthocyanin. Figure 2 shows this pathway illustrated for pelargonidin. In the case of purple pods cyanidin and delphinidin are present, and they differ from pelargonidin by the addition of one or two hydroxyl groups, respectively, to the aromatic ring B.

In pea, apart from *A* (see above), most of the pigmentation loci investigated are involved with colour modification rather than its initial production (Statham *et al.*,1972) and so are unlikely to encode those enzymes involved in anthocyanin biosynthesis. *Am*, *Ar*, and *Ce* have quantitative effects and so may encode regulatory proteins, although they may represent down mutations of structural genes. *B* appears to encode the enzyme which hyroxylates the B-ring of anthocyanins (so producing cyanidin and delphinidin rather than pelargonidin) and *Cr* encodes the enzyme which subsequently methylates these hydroxyl groups in flowers. The *Cgf* phenotype results in a diversion of the anthocyanin biosynthesis pathway, petals accumulate flavones (Statham & Murfet, 1974) which are formed by the desaturation of flavanones (Heller & Forkmann, 1988). Whether this phenotype is due to a sequestration of flavanone by a flavone synthase or to a blockage of the flavanone 3-hydroxylase step is unknown.

Regulation of the anthocyanin biosynthesis pathway

In addition to the structural genes isolated from the anthocyanin biosynthesis pathway, regulatory genes have also been characterised, again primarily from *Antirrhinum*, maize and *Petunia* (Dooner *et al.*,1991). The regulatory genes appear to act in a coordinate manner controlling the expression of more than one of the steps in the anthocyanin biosynthesis pathway, (Dooner, 1983, Ludwig *et al.*,1989, Goodrich *et al.*,1992). The maize regulatory genes comprise a small multigene family each member causing pigmentation of different tissues or groups of tissues (Ludwig & Wessler,1990) and a similar situation may occur in *Antirrhinum* with a member of the *delila* gene family (Goodrich *et al.*,1992).

Both the maize *R* gene family and the *Antirrhinum delila* gene encode proteins with homology to each other and to helix-loop-helix transcription factors isolated from other eukaryotes (Ludwig & Wessler, 1990, Goodrich *et al.*,1992). The implication is that these regulatory genes encode DNA-binding transcriptional activators but this remains to be proven. In pea the regulatory loci and *2* have been shown to control chalcone synthase production in petals (Harker *et al.*,1990) but as yet these loci have not been characterised at a molecular level, nor is it known whether they regulate other genes from the pathway.

Instability in the purple podded phenotype

Whilst growing the commercial "Purple Podded" pea line it was noticed that this phenotype was unstable in some plants. Pods appeared with purple spots to varying density (fig. 3), rather than the overall purple of the phehotype. When seeds from these spotted podded plants were grown, some progeny plants retained the spotted pod characteristic, and some plants produced totally green pods. No reversion from spotted or

Fig. 3 A purple spotted pod on a mutant plant from the Purple Podded pea line.

green pod bearing plants to purple pod bearing plants was seen. The pigmentation of the flowers and maculum rings at the axil of the stipule remained unaffected in these plants (Bown,D. & Gatehouse,JA., unpublished results). This instability of the phenotype had previously been noted by Lamprecht (1941), who described four alleles of the *Pur* locus; *Pur-Pur*^{a}*-Pur*^{b}*-pur* with mutation occurring at varying frequencies in the stages between *Pur>pur*.

The nature of this instability is very similar to colour mutations known to be the result of the action of transposable elements. Such mutable alleles are found in a wide variety of plant species and affect

many characteristics (Nevers *et al.*,1986), the most obvious being those which affect colour production. The species in which transposble elements and their effects on anthocyanin biosynthis have been most widely studied are *Zea mays* (maize) and *Antirrhinum majus* (Snapdragon). Many of the structural and regulatory genes from this pathway have been found to be mutated in various lines leading to variegated phenotypes (Coen & Carpenter,1986, Coen *et al.*,1988, Wessler,1988). For example, in maize, instability at the *A1* locus leads to coloured sectors on the seed coat aleurone layer and recessive mutations to total lack of colour. The *A1* locus has been shown to encode dihydroflavonol reductase (fig. 2) and the instability to be due to the action of transposable elements (Reddy *et al.*,1987, Schwartz-Sommer *et al.*,1987, O'Reilly *et al.*,1985).

Transposable elements and their possible role in the unstable purple podded phenotype

Transposable elements are sections of DNA which are able to move within the genome. This movement gives rise to unstable genes with mutation rates much greater than that associated with spontaneous mutation (10^5-10^6 generations, Coen & Carpenter, 1986) in the otherwise very stable genome. Two classes of transposable elements occur; class one elements transpose via reverse transcription of an RNA intermediate, whereas class two elements transpose directly from DNA to DNA (Finnegan, 1989). Most of the elements characterised in plants belong to the second class, although class one type elements have been found in plants, including pea (Lee *et al.*,1990).

The class two elements have terminal inverted repeats which range from 6 to around 200bp in length and, upon insertion, cause a duplication of the host sequence resulting in direct repetition of 3-9bp either side of the element (Vodkin, 1989). Similarity in the sequence of

the terminal repeats and in the length of target site duplications can be seen between elements isolated from different plant species, and it seems that transposon families span a wide range of host species. The internal structure of transposable elements is less well characterised (Vodkin, 1989, Weil & Wessler, 1990); it varies within and between families and determines whether elements are autonomous (can transpose themselves) or non-autonomous (require gene products from another, autonomous, element to transpose). The autonomous element must contain a transposase together with a factor which recognises the terminal repeats whereas the non-autonomous elements have lost this function but retain the terminal repeats. Apart from this, transposons may contain gene products that suppress the transcription of host genes whilst boosting their own.

As stated above, integration of transposable elements leads to the duplication of host target sequence. Subsequent excision of the element rarely restores the original wild-type sequence, as all or part of the duplication remains (Wiel & Wessler, 1990). Whilst this additional sequence may not be deleterious in flanking or intron sequence, in coding sequence a resultant frame shift would cause premature termination or nonsense translation. The presence of a transposable element may also affect gene expression not only by its physical separation of the component parts of the gene (although elements may be spliced from mRNA) but by the insertion of new regulatory sequences proximal to the existing gene promoters.

Transposable elements, once isolated, may be utilised to clone genes whose products are unknown but whose expression is altered by the insertion of a transposable element. Depending on the frequency of occurrence of the elements within a genome, the presence of a transposable element may be used as a "tag" to select a gene containing

the element. Such transposon tagging strategies have been successfully used to isolate structural and regulatory genes from the anthocyanin biosynthesis pathways of maize and *Antirrhinum* (Fedoroff *et al.*,1984, O'Reilly *et al.*,1985, Cone *et al.*, 1986, Goodrich *et al.*,1992).

In the case of the variable purple podded phenotype the most obvious explanation would be that a transposable element was inserted in such a way as to block the expression of a structural gene in the anthocyanin biosynthesis pathway. Excision of the element from the locus in somatic pod tissue leads to restoration of gene function in some cells and their progeny, giving rise to purple spots. However this model fails to explain the observation that the phenotype mutates from *Pur* (purple podded) via intermediates to *pur* (green podded) but not vice-versa (see above). Stable green podded mutants may be the result of imprecise excision of the element in germ-cells, but in that case, why are stable purple revertants not encountered? The model predicts that restoration of gene function should occur in the same way as it does in somatic tissue. The nature of this phenotypic instability can only satisfactorally be answered by molecular analysis of this locus in wild type, unstable and stable green tissue.

Transposable elements in pea and other legumes

Elements carrying the features of transposable elements have been isolated from legumes but no active systems have yet been characterised. In pea, legumin gene *legC* contains a 2.5 kbp insertion in the 5' flanking sequence with respect to *legB*. The termini of this insertion form inverted repeats with sequence homology to the termini of transposable elements *En/Spm* from maize and *Tam1* and *Tam2* from *Antirrhinum*, and these repeats are flanked by a 3bp duplication of the *legB* sequence (Shirsat, 1988). The wrinkled seed characteristic in pea

was found to be the result of the insertion of a 0.8kbp element in the 3' coding sequence of the starch-branching enzyme. Again this insertion has inverted terminal repeats, this time homologous to *Ac/Ds* elements fom maize and *Tam3* from *Antirrhinum*, and is flanked by 8bp repeats of the target gene sequence (Bhattacharyya *et al.*,1990).

A retrotransposon-like element with homology to *copia* elements from *Drosophila* has also been isolated from pea between legumin genes *legJ* and *legK* and there are indications that this element may be active (Lee *et al.*,1990). In soybean, the lectinless phenotype was found to be due to a 3.5kbp insertion within lectin encoding sequence with similar characteristics to those of the pea *legC* insertion (Vodkin *et al.*,1983). Further work has revealed the presence of a family of such elements, those isolated being deletion mutants of a putative 16kb progenitor element (Rhodes & Vodkin, 1988). Hybridisation studies with all these legume elements have shown them to be repeated many times in the genome.

In addition to the unstable purple-podded locus, various other unstable phenotypes have been described in pea. A single plant with a mutation causing similar effects in flowers to those seen in pods was propagated, although no material now remains. Flowers from an originally purple line appeared white with varying degrees of purple spotting. This phenotype appears to have been due to instability at the *A* locus and therefore, unlike the purple-podded mutation, pigmentation in other tisues, maculum ring and seed coat, was also affected (De Haan, 1930). Instability has also been noted in the distribution of chlorophyll, pod schlerenchymus tissue and testa colour (De Haan, 1930, Lamprecht, 1941). While these are all obvious, non-lethal mutations, presumably many others occur which are unnoticed because they affect one of multiple alleles or whose general effect is obvious only as poor growth.

Specific aims of the project

As part of the ongoing study of pea seed storage protein within this department, and in collaboration with the John Innes Institute, genes encoding various storage proteins had been isolated. The initial aim was to determine the nucleotide sequence of these genes and to compare them with other members of their gene families in pea and other legumes. Secondly, the expression of these genes was to be investigated, the organisation of the gene families they belong to clarified, and their specific products identified.

There had been so little study of root protein that it was decided to investigate and characterise proteins that accumulate to high levels in the pea root and the genes which encode them. Some legume species exist with tuberous roots which function as storage organs during periods of dormancy, for example, *Phaseolus coccineus* (runner bean), *Psophocarpus tetragonolobus* (winged bean), *Vigna vexillata* (cowpea), *Sphenostylis stenocarpa* and *Pachyrrhizus erosus* (yam beans). These presumably contain storage proteins, and it would be interesting to determine whether legume species which do not form tuberous roots accumulate any such storage proteins. Therefore, with root tissue the primary objective was to determine whether there were any abundant proteins present, and if so, to purify and characterise them. With this information, genes encoding these proteins could be isolated using oligonucleotide probes generated from aa sequence data.

In order to investigate the purple-podded phenotype and the instability within it, a differential screen was to be conducted between cDNA from purple and green podded pea lines. cDNAs isolated by this method could then be used (together with cDNAs encoding enzymes from the anthocyanin biosynthesis pathway) to try to determine the cause of the instability in the colour production by comparing their expression in

tissue from mutant and normal pod coloured plants. These cDNAs could then be used as probes to isolate genes from a genomic library from the purple-podded line.

CHAPTER TWO: MATERIALS

Chemical and equipment suppliers

Unless otherwise stated, chemicals were of analytical grade and supplied by BDH-Merck Ltd., Lutterworth, Leics., Medicell dialysis tubing was also obtained from BDH.

The following reagents were supplied by Boehringer Mannheim UK Ltd., Lewes, E. Sussex:

cDNA synthesis kits, Caesium Chloride, Glycogen, MI3 mp18 & 19 DNA, Oligo d(T) cellulose, Restriction & DNA modifying enzymes.

The following reagents were supplied by Sigma Chemical Co. Ltd., Poole, Dorset:

Acrylamides, Antibiotics, BSA, CNBr, DTT, Ethidium Bromide, Herring sperm DNA, Iodoacetamide, Leupeptin, PIPES, PMSF, Polyadenylic acid, Pronase P, Proteinase K, RNase, tRNA.

The following reagents were supplied by Gibco BRL, Life Technologies Ltd., Paisley, Scotland:

Agarose, DH5α competent cells, HGT agarose, Manual DNA sequencing kits - including JM101 bacteria.

The following reagents were supplied by Pharmacia Biosystems Ltd., Milton Keynes:

cDNA synthesis kits, *Eco* RI - blunt end adaptors, Ficoll, Minigel apparatus, Random hexanucleotide DNA oligo, Sephacryl S-200, Sephadex G-50, Sepharose CL-4B, S1 nuclease.

The following reagents were supplied by Promega Ltd., Southampton:

λGEM-12 *Xho* half site arms cloning system including KW251 bacteria, Packagene packaging extract, Protoclone λgt10 including C600 bacteria, *Taq* polymerase.

The following reagents were supplied by Northumbria Biologicals Ltd., Cramlington, Northumberland:

DNA size markers, IPTG, Plasmids pUC18 & 19, Restriction and DNA modifying enzymes, X-Gal.

Radiochemicals were supplied by Amersham International plc., Aylesbury, Bucks.

National Diagnostics "Ecoscint", scintillation fluid, was supplied by B.S.& S (Scotland Ltd.), Edinburgh.

Nitrocellulose, Schleicher & Schuell grade BA-85, was supplied by Anderman & Co. Ltd., Kingston-upon Thames, Surrey.

3MM filter paper was supplied by Whatman Labsales Ltd., Maidstone, Kent.

X-ray film, Fuji-RX, was supplied by Fuji Photo Film (UK) Ltd., London.

X-ray cassettes and ATTO protein gel and western blotting apparatus were supplied by Genetic research instrumentation Ltd., Dunmow, Essex.

Intensifying screens, Cronex lightning-plus were supplied by Dupont UK Ltd., Stevenage, Herts.

Developer, Ilford phenisol, was supplied by Ilford Ltd., Mobberly, Ches.

Fixer, Kodak Unifix, was supplied by Phase separations Ltd., Deeside, Clwyd.

Perfect match enhancer was supplied by Stratagene Ltd., Cambridge.

Quiagen DNA purification kits and PCR heating block were supplied by Hybaid Ltd., Teddington, Middlesex.

Immun-Blot western blotting kit was supplied by Biorad Laboratories Ltd., Hemel Hempstead, Herts.

DNA synthesiser and sequencers were supplied by Applied Biosystems Inc., Warrington, Ches.

Microtitre plates and disposable pipettes and petri dishes were supplied by Bibby Sterilin Ltd., Stone, Staffs.

Bacto-agar was supplied by Difco Ltd., W. Molesey, Surrey.

Oxoid Yeast extract was supplied by Unipath Ltd., Basingstoke, Hants.

Bacto-tryptone was supplied by Becton Dickinson, Cowley Oxon.

Phostrogen was obtained from Phostrogen Ltd., Corwen, Clywd.

Frequently used buffers, media and other solutions

TAE buffer	40mM Tris-acetate pH 7.7 10mM EDTA
Denaturing solution	1.5M NaCl 0.5M NaOH 1mM EDTA
Neutralising Buffer	3M NaCl 0.5M Tris-HCl pH 7.0 1mM EDTA
20xSSC	3M NaCL 0.3M Tri-sodium citrate pH 7.0 with HCl
Denhardt's solution (1x)	0.02% Ficoll 400 0.02% Polyvinyl-pyrrolidone 0.02% BSA
TE buffer	50mM Tris-HCl pH 7.5 10mM EDTA
Phage buffer	20mM Tris-HCl pH 7.4 0.1M NaCl 10mM $MgSO_4$
LB medium, per litre	10g Bacto-tryptone 5g yeast extract 5g NaCl pH 7.5 with NaOH

LB agar, as above + 15g l^{-1} Bacto-agar

TB top agar, per litre	10g Bacto-tryptone 5g NaCl 8g Bacto-agar
YT medium, per litre	8g Bacto-tryptone 5g Yeast extract 5g NaCl

YT agar, as above + 15g l^{-1} Bacto-agar

YT top agar, YT medium + 6g l^{-1} Bacto-agar

2xYT medium, per litre	
	16g Bacto-tryptone
	10g Yeast extract
	5g NaCl

Ampicilin was added to media and plates as required, at 50μg ml^{-1} final concentration

X-gal was added to plates as required, at 40μg ml^{-1} final concentration

SOC medium	
	2% Bacto-tryptone
	0.5% Yeast extract
	10mM NaCl
	2.5mM KCl
	10mM $MgCl_2$
	10mM $MgSO_4$
	20mM Glucose

Phenol was redistilled, saturated with TE buffer and stored frozen under N_2.

Formamide was deionized by stirring for 2hr with amberlite MB1 resin, then filtered and stored frozen.

Treatment of glassware, plasticware and solutions

Glassware, plasticware and solutions for use with DNA and bacteria were sterilised by autoclaving for 20min at 120°C. Glassware was siliconised before use with nucleic acid. Glassware for use with RNA was baked overnight at 170°C. Solutions to be used with RNA were incubated with diethylpyrocarbonate at 0.1% final concentration, overnight at room temperature, before autoclaving as above.

Plant material

Garden pea (*Pisum sativum* L.) seed material was obtained from the following suppliers: Feltham First (FF) and Purple podded (PP) cvs. from Sutton Seeds Ltd., Torquay, Devon; Dark skinned perfection (DSP) from Samuel Dobie & Son Ltd., Llangollen, Clywd; Birte (line 2791) from The Nordic Gene Bank, Alnarp, Sweden.

Seeds were germinated in the dark, in a spray room for 4-5 days. Plants were then grown in a growth cabinet, hydroponically in Phostrogen

solution prepared according to the manufacturer's instructions. Growth cabinet conditions were: temperature; 25°C 06.30-20.30hr, 17°C 20.30-06.30hr; humidity, 70% relative humidity; flourescent lighting on 06.00-18.00hr, doubled intensity 08.00-16.00hr; incandescent lighting on 15.00-18.00hr. Phostrogen solution was changed regularly and flowers were tagged when fully open.

Bacterial strains

The genotypes of the bacterial strains used are:

JM101; *sup*E, *thi*, Δ(*lac pro*A,B)/F', *tra*D36, *pro*A,B, (r_K^+, m_K^+), *lac*I^qZ M15.

DH5α; F-, *end*A1, *hsd*R17 (r_K^-, m_K^+), *sup*E44, *thi*-1, λ-, *rec*A1, *gyr*A96, *rel*A1, Δ(*arg*F-*lac*zya)U169, ϕ80d*lac*ZΔM15.

C600*Hfl*; *hsd*R^-, *hsd*M^+, *sup*E44, *lac*Y1, *ton*A21, *hfl*A150, (*chr*:Tn10).

KW251; F-, *sup*E44, *sup*F58, *gal*K2, *gal*T22, *met*B1, *hsd*R2, *mcr*B1, *mcr*A-, *arg*A81:*Tn*10, *rec*D1014

Plasmid and phage DNA restrictions

Restriction of plasmid and phage DNA was performed at 37°C for 2hrs using buffer supplied with the restriction enzymes. A restriction enzyme to DNA ratio of at least 1 unit μg^{-1} was used and the final dilution of enzyme was at least 10x.

Phenol extraction and ethanol precipitation of DNA

Unless otherwise stated, DNA was phenol extracted as follows, vortexing for 15sec and spinning in a microcentrifuge for 3min at 12,000g at each stage. The volume was adjusted up to 200μl with TE buffer if required, and extracted with 200μl phenol. After vortexing and centrifuging the aqueous phase was taken and retained. The phenolic phase was extracted with 100μl TE buffer, vortexed and spun. This aqueous phase was added to the previous aqueous phase and extracted with 300μl chloroform/isoamyl alcohol (24:1), vortexing and spinning as above. The aqueous phase was removed, reextracted as above and the final aqueous phase removed and ethanol precipitated.

Unless otherwise stated, DNA was ethanol precipitated using 1μl (20μg) glycogen carrier, adding 1/25 volumes of 5M ammonium acetate 0.25M $MgCl_2$ pH 5.2 and two volumes of ethanol cooled to -20°C. This was mixed and placed at -20°C for at least 1hr followed by centrifugation at 4°C, 12,000g for 20 mins. The ethanol was carefully poured from the tube and replaced by 70% ethanol at -20°C. Tubes were inverted to mix, spun as above for 5min and the liquid poured off. The DNA was dried under vacuum for 5min and resuspended in H_2O.

Agarose gel electrophoresis of DNA

Electrophoresis of DNA was carried out in a Pharmacia GNA-100 minigel apparatus or on 15.5x18.5x0.6cm gels in tanks holding 2.1l of buffer as described in Maniatis *et al.*(1982). TAE buffer was used, gels and running buffer contained 1μg ml^{-1} ethidium bromide. 0.2-0.5 volumes of dye mix (10mM Tris-HCl, 10mM EDTA, pH8.0, 1mg ml^{-1} fast orange G., 30% glycerol) were added to samples before loading. Commercially prepared restriction digests of λ DNA were run as size markers. A voltage of 2-10 volts cm^{-1} was applied and gels were photographed using transmitted UV light at 300nm.

Isolation of DNA from agarose gel

Slices of agarose gel containing DNA were placed in prepared dialysis tubing (Maniatis *et al.*,1982) with the minimum volume of TAE buffer, all air excluded and the ends clipped shut. The tubing was placed in a Pharmacia GNA-100 minigel apparatus, perpendicular to the field and covered with TAE buffer. DNA was electrophoresed from the gel for 15min at 50V. After checking under UV that the DNA had migrated to the edge of the dialysis membrane, the polarity was reversed for 15sec. Using a micropipette, buffer containing the DNA was removed from around the gel and the tubing was then rinsed with a small (50-100μl) volume of TAE. The DNA was then puified by phenol extraction and ethanol precipitation.

Determination of nucleic acid concentration

Unless otherwise stated, nucleic acid concentration was determined spectrophotometrically on a Pye Unicam SP-800 dual beam spectrophotometer. A 1mg ml^{-1} solution of DNA was assumed to give an OD_{260} of 20 and a 1mg ml^{-1} solution of RNA an OD_{260} of 25. When a very

accurate determination of DNA concentration was required, for genomic blots, this was performed on plasmid and genomic DNA preparations using diaminobenzoic acid by the method of Thomas and Farquhar (1978).

General subloning procedure

Inserts to be subcloned were restricted, phenol and chloroform/isoamyl alcohol extracted, and ethanol precipitated. Vector, usually M13 mp18 or 19 for sequencing and plasmid pUC18 or 19 for DNA required double stranded (to be used for labelling or further subcloning), was also processed in the same way. Ligations containing equimolar amount of insert and vector (using 0.1-0.2μg vector) were set up in 10μl volume and incubated overnight at 15°C. 1unit of T_4 DNA ligase was used in each reaction with buffer supplied with the enzyme.

Phosphatase treatment of vector DNA

Vector DNA, restricted as above was phosphatase treated using the method in Maniatis *et al.* (1982). Before being used in cloning, DNA was phenol extracted and ethanol precipitated.

Plasmid transformation of competent cells

DH5α competent cells were transformed following the supplier's instructions: an aliquot of ligated plasmid (2-10ng) was gently mixed with freshly thawed cells and incubated on ice for 30min. Cells were heat shocked for 90sec at 45°C and then four volumes of SOC medium added and the cells incubated at 37°C for 1hr to express ampicilin resistance. Aliquots were then spread over YT amp Xgal plates and incubated overnight at 37°C.

Minipreparation of plasmid DNA for restriction analysis

Single colonies from agar plates, or loopfulls from glycerol preserved cells, were innoculated into 10ml aliquots of YTamp and incubated overnight at 37°C on a rotating wheel. Plasmid DNA was prepared from these cultures by a scaled up version of the method of Birnboim and Doly (1979) using the solutions described therein.

Bacteria were harvested by centrifugation, resuspended in 200μl solution I in a microcentrifuge tube and kept on ice for 30 mins. 600μl of solution II were added, the tube kept on ice for 5min and 450μl of solution III added. The tube was inverted to mix and incubated on ice for 1hr. To 1.1ml of the clear supernatent from a 5min, 12,000g centrifugation was added 500μl of -20°C isopropanol and the contents mixed and placed at -20°C for 30min.

The precipitate from a 2min, 12,000g centrifugation was resuspended in 400μl 0.1M sodium acetate, 50mM Tris-HCl pH6.0 and reprecipitated with 1ml of -20°C ethanol for 10 min at -20°C, centrifuging as above. This pellet was resuspended and reprecipitated as in the last steps and finally dried for 5min under vacuum and resuspended in 100μl H_2O. 5μl aliquots were used for restriction analysis.

Minipreparation of plasmid DNA for sequencing

Bacteria containing plasmid DNA was grown overnight as above and harvested by centrifugation. DNA was purified for sequencing by a further modification of the Birnboim and Doly method (above). Cells were resuspended in 200μl of solution I, kept on ice for 30min, 400μl of solution II added and the tube mixed by inversion and kept on ice for 5min. 300μl of acid potassium acetate (600μl 5M potassium acetate + 115μl glacial acetic acid + 285μl H_2O) were mixed in by invertion and the mixture placed on ice for 30 min.

The tube was then centrifuged for 30min at 12,000g and 700μl of the clear supernatent removed. To this was added 2μl of 10 mg ml^{-1} RNase (DNase free) and the tube incubated at 37°C for 20min. The solution was then extracted once with phenol/chloroform/isoamyl alcohol (25:24:1) and once with chloroform/isoamyl alcohol (24:1), vortexing for 15sec and centrifuging for 3min at both stages. DNA was then ethanol precipitated as usual.

The DNA pellet was resuspended in 16.8μl H_2O, 3.2μl of 5M NaCl were then mixed in, followed by 20μl of 13% polyethylene glycol 8000 and the contents mixed and stood on ice for 20min. DNA was recovered by centrifugation for 10min at 12,000g and the supernatent removed by micropipette. 1ml of -20°C 70% ethanol was added, the tube spun for a further 5min and the supernatent decanted. The precipitate was dried under vacuum and resuspended in 15μl H_2O. 1μl aliquots were restricted and electrophoresed with known amounts of standard DNA to estimate the concentration prior to sequencing.

Transformation of competent cells with M13 vector

Commercially prepared frozen DH5α cells were transformed with M13 following the supplier's protocol. Aliquots of diluted ligation (2-10ng DNA) were mixed gently with freshly thawed cells and incubated on ice for 40min. These were then heat shocked at 45°C for 2min, cooled and mixed with 200μl of freshly grown JM101 cells, 10μl of 0.1M IPTG, 50μl 2% X-Gal in dimethylformamide and 3ml YT top agar at 45°C. This mixture was immediately poured onto YT agar plates, allowed to set and incubated overnight at 37°C.

Minipreparation of M13 DNA

M13 DNA for sequencing was minipreped using a method supplied by Applied Biosystems (model 373A user's manual). Recombinant (clear) plaques were picked off using a sterile cocktail stick, into 1.8ml 2xYT broth containing 1μl ml^{-1} of a recently grown culture of JM101 cells, and incubated overnight at 37°C. Cultures were transferred to microcentrifuge tubes and centrifuged at 12,000g for 5min. The supernatants were transferred to fresh tubes and the pellets stored at 4°C. The supernatants were then centrifuged for a further 10min, 1.25ml removed to fresh tubes, 125μl of 5M NaCl and 125μl of polyethylene glycol mixed in and these incubated at room temperature for 20min.

Phage was recovered by spinning for 5min, pouring off the supernatant, respinning for 10sec and then removing the last traces of supernatant with a micropipette. The precipitate was resuspended in 200μl TE buffer and extracted twice with phenol and once with chloroform/isoamyl alcohol (24:1) using 200μl each time, vortexing for 15sec and centrifuging at 12,000g for 3min. The final aqueous phase was then extracted twice with H_2O saturated ether, vortexing and spinning as above. 20μl of 3M sodium acetete pH4.8 and 500μl of ethanol at -20°C were added and the contents mixed and incubated for at least 1hr at -20°C.

DNA was collected by centrifugation at 12,000g for 15min, the pellet rinsed with 1ml -20°C 70% ethanol - spinning for 5min, dried under vacuum for 5min and resuspended in 20μl TE buffer. 1μl aliquots were electrophoresed on agarose gel alongside a known quantity of untransformed M13 to assess concentration and size of insert.

Complimentarity test on M13 transformants

To assess which orientation inserts were with respect to each other in M13, complimentarity or "C" tests (Messing, 1983) were performed. 1μl aliquots of minipreped DNA from the clones to be tested were mixed and to them added 12μl TE buffer, 7.5μl 1M NaCl, and 5μl dye mix (3% SDS, 0.1% bromophenol blue, 60% deionized formamide, 25mM EDTA). This was mixed, incubated at 65°C for 1hr, and then electrophoresed on a 0.7% agarose gel. When inserts of opposite orientation are present their sequence is complimentary and the hybridised molecules have a lower mobility through the gel than single circular transformed M13.

Manual DNA sequencing

Manual DNA sequencing was carried out by the dideoxy chain termination method (Sanger *et al.*,1977), using ^{35}S dATP and BRL M13 sequencing kits, by Mr.P. Preston and Ms.J. Bryden. Where suitable restriction sites were unavailable, synthetic oligo-nucleotide primers were synthesised against determined sequence and these used to prime the sequencing reactions in place of the M13 primers.

Automated DNA sequencing

Automated DNA sequencing was performed on an ABI model 373A DNA sequencer, by Ms.J. Bryden, using ABI reagents in a modification of the dideoxy chain termination method with *Taq* polymerase (see 373A user's manual).

Extraction of RNA from plant tissues

Plant tissue was harvested, immediately frozen in liquid N_2 and stored at -80°C until required. RNA was extracted using the hot SDS method (Hall *et al.*,1978) and stored at -80°C.

Selection of poly(A) enriched RNA

Poly(A) enriched RNA was selected on oligo d(T) cellulose columns from total RNA prepared as above, following the method in Maniatis *et al.* (1982) and stored under liquid N_2.

Formamide RNA gels

RNA was electrophoresed in agarose gels containing formaldehyde, after denaturation with formamide (Miller, 1987). 15.5x18.5cm gels were prepared by boiling 1.4g HGT agarose in 63ml H_2O until dissolved, this was cooled to 70°C, 9.3ml 0.5M MOPS, 10mM EDTA pH7.0 and 17ml 37% formaldehyde mixed in and the gel poured.

RNA samples in <3μl volume were denatured by adding 4.4μl buffer A (see below), 11.6μl formamide/formaldehyde (see below) and heating to 70°C for 10min. Samples were cooled on ice before adding 1.5μl gel loading buffer (see below) and loading. Gels were run at 100V in tanks holding 2.1l of 50mM MOPS 1mM EDTA pH7.0, which was constantly stirred and circulated. Pea and *E.coli* ribosomal RNA and cowpea chlorotic mottle virus RNA were used as standard size markers.

After electrophoresis, the portion of the gel to be blotted was placed directly on the blotting apparatus (see gel blotting) and the rest stained for 5min in 5μg ml^{-1} ethidium bromide, destained overnight in H_2O, and photographed using transmitted UV light at 300nm.

Buffer A:	294μl 0.5M MOPS 10mM EDTA pH 7.0 706μl H_2O
Formamide/formaldehyde:	89μl 37% formaldehyde 250μl deionized formamide
Gel loading buffer:	2μl 37% formaldehyde 5μl deionized formamide 7μl Gel dyes
Gel dyes:	322μl Buffer A 5mg bromophenol blue 400mg sucrose

Blotting of agarose gels onto nitrocellulose filters

Gels containing nucleic acid were blotted onto nitrocellulose filters basically as described in Maniatis *et al.* (1982). Gels containing high Mr (>10kbp) DNA were first soaked in 0.25M HCl for 2x15min, gels containing DNA were then soaked for 2x30min in denaturing solution followed by 2x30min in neutralising buffer. All gels were then placed on the apparatus as described, except that nappy liners were used in place of paper towels and 20xSSC was used as the transfer buffer. After blotting, the position of the wells was marked on the filter with ink and nucleic acid was bound to the nitrocellulose by baking at 80°C under vacuum for 2 hr.

Random primed labelling of DNA

DNA for labelling was prepared by restriction and electroelution. DNA was labelled with α^{32}P dCTP (100ng quantities labelled with 50μCi of >400Ci mmol^{-1}) by the random primed method (Feinberg & Vogelstein,1984), incubating overnight, to a specific activity >10^7 cpm μg^{-1}. The reaction was stopped with 1/20 volumes of 20% SDS and DNA was separated from unincorporated label by gel filtration through a 5ml column of Sephadex G-50 (run in 0.15M NaCl,10mM EDTA,0.1%SDS,50mM Tris-HCl pH7.5). The excluded peak (judged with gieger monitor) was collected and samples taken for counting with "Ecoscint" scintillant. Before use, probes were boiled for 5min to render the DNA single-stranded.

Oligonucleotide synthesis

Oligo deoxynucleotides were synthesised on an ABI model 381 DNA synthesiser by Mr. JS.Gilroy.

5' End Labelling of DNA

If present, 5'-terminal phosphate groups were removed from DNA using calf intestinal alkaline phosphatase (Maniatis *et al.*,1982). DNA was then phenol and chloroform extracted and ethanol precipitated (using sodium acetate pH4.8 to 0.3M in place of ammonium acetate). DNA was end labelled with T4 polynucleotide kinase using ^{32}P γATP (6000 Ci mmol^{-1}) as per Maniatis *et al.* (1982). The amount of radioactivity incorporated into DNA was determined by TCA precipitation and scintillation counting (Maniatis *et al.*,1982).

Hybridisation of DNA probes to filters containing DNA

Nitrocellulose filters containing DNA were processed in sealed polythene bags in a shaking water bath at 65°C with solutions preheated to this temperature. 100ml of solution per 20x20cm filter was used except for hybridisation when the volume used was halved.

Filters were prehybridised in 5xSSC,5xDenhardt's solution,100-200μg ml^{-1} denatured herring sperm DNA (prepared as in Maniatis *et al.*,1982) for at least 2hr. This solution was then replaced with 5xSSC,1-2xDenhardt's solution, 100μg ml^{-1} denatured herring sperm DNA, labelled probe was added and hybridisation allowed to proceed overnight or longer.

Hybridisation solution was then removed and the filters washed sequentially with the following solutions for 30min each until the required stringency was reached; 2xSSC, 1xSSC twice, 0.1xSSC twice. Filters were removed from the polythene bags, blotted dry and exposed to X Ray film.

Hybridisation of RNA containing filters with labelled DNA

Nitrocellulose filters containing RNA were processed in sealed polythene bags in a shaking water bath at 42°C, with solutions preheated to this temperature. 100ml of solution per 20x20cm filter was used, except for hybridisation, when the volume was halved. When probes contained poly(A) tails (or could possibly do so) 100μg ml^{-1} poly(A) was added to the prehybridisation and hybridisation solutions.

Prehybridisation was in 50% deionised formamide, 5xDenhardt's solution, 100μg ml^{-1} denatured herring sperm DNA (prepared as in Maniatis *et al.*,1982), 0.1%SDS, for at least 2hr. This solution was replaced with 50% deionized formamide, 2xDenhardt's solution, 100μg ml^{-1} denatured herring sperm DNA, 0.1% SDS, 5xSSC, and the probe added. Hybridisation was allowed to proceed at least overnight before washing sequentially with the following solutions, for 30min each until the required stringency was reached; (2xSSC 0.1%SDS) twice, (0.1xSSC 0.1%SDS) twice. Higher stringency was acheived by performing the final washes at 50°C. Finally, filters were removed from the bags, blotted dry and exposed to X-ray film.

Hydridisation of oligonucleotide probes to northern blots

Northern blots were probed with end labelled oligonucleotides using the same methods as for longer probes except that the solutions used were taken from Woods (1984). Prehybridisation was performed for at least 2hr at the temperature specified (see results) in 6xSSC, 1xDenhardts solution, 0.5% SDS, 100μg ml^{-1} herring sperm DNA, 0.05% sodium pyrophosphate. Hybridisation was performed overnight at the same temperature in 6xSSC, 1xDenhardts solution, 20μg ml^{-1} tRNA, 0.05% sodium pyrophosphate. Washing was as specified (results section), initially at 5xSSC, 1%SDS and subsequently at 3xSSC and 1xSSC, 1%SDS all for 2x30min.

Exposure of filters to X-Ray film - Autoradiography

Nitrocellulose filters probed with ^{32}P labelled DNA were secured to card, the origin of gels, or surround of filters bearing bacteria, marked with an asymetrical pattern of radio-active ink (writing ink with a small quantity of ^{32}P) and this assembly placed in a polythene bag. A pre-flashed X-Ray film was placed between the filter and an intensifying screen within a cassette and exposed (for periods longer than 6hr, at -80°C). Films were developed following the manufacturer's instructions for the developer and fixer used.

Construction of pea pod cDNA library in plasmid vector

cDNA was made from 5μg of poly(A) enriched RNA from 5 daf. PP. pea pods using a Pharmacia cDNA synthesis kit, following the kit instructions and using the reagents supplied. The RNA was heat denatured for 10 min at 65°C in a total volume of 20μl, then chilled on ice. First-strand cDNA was synthesised using an oligo $d(T)_{12-18}$ primer and Moloney murine leukemia virus reverse transcriptase. Second-strand cDNA was produced using DNA polymerase I, after nicking the RNA with RNase H, Klenow fragment of DNA polymerase was then used to produce blunt-ended molecules. Enzymes were finally denatured by heating at 65°C for 10 min.

The cDNA was extracted with phenol/chloroform/isoamyl alcohol (25:24:1) and then purified on a Sephacryl S-300 spun column equilibrated with ligation buffer. Blunt end to *Eco* RI adaptors were ligated to the cDNA overnight at 12°C, heated at 65°C for 10 mins, phenol/chloroform/isoamyl alcohol extracted and purified from non-ligated linkers on another spun column.

Trial ligations were performed using 1/25, 1/50, 1/75 and 0 of the linkered cDNA, each ligated to 0.1μg of dephosphorylated, *Eco* RI digested pUC19 vector, in a total volume of 36μl, overnight at 12°C. All

of each ligation was then used to transform 200μl of frozen competent DH5α cells and 10 and 100μl aliquots (out of 1ml), plated onto YT amp Xgal agar and incubated overnight at 37°C.

The remaining test ligations were then plated in 200μl aliquots, grown overnight and the total library, representing $2x10^4$ colonies, scraped off the plates with YT medium and preserved by mixing with an equal volume of 80% glycerol and storing at -80°C. cDNA not used for cloning was stored at -20°C for subsequent random prime labelling.

Preservation of plasmid cDNA library in microtitre plates

Single colonies of transformed bacteria were innoculated into microtitre plate wells containing 130μl of 2xYTamp and these were incubated overnight at 37°C. 130μl of 80% glycerol was added to each well, mixed thoroughly and 130μl transferred to the same position in a second plate. Both plates were then stored at -20°C, one plate being retained at this temperature as a master.

Amplification of bacterial colonies on nitrocellulose filters

Nitrocellulose filters containing bacterial colonies originally grown overnight on YTamp agar plates, were transferred to plates containing YT agar supplemented with 150μg ml^{-1} chloramphenicol and incubated for 24hr at 37°C.

Lysis of bacteria on nitrocellulose filters

Bacteria on nitrocellulose filters were lysed by placing the filters in turn, for 5min each, onto blotting paper soaked in the following solutions, drying on blotting paper between each step; 10% SDS, Denaturing solution, Neutralising buffer, 2xSSC. DNA was then bound to the filters by baking at 80°C under vacuum for 2hr.

Preparation of pod cDNA for labelling

1μg of poly(A) enriched 5daf. FF. pod RNA was used to make cDNA, using a Boehringer kit as described in the manufacturer's instructions. First-strand synthesis utilised an oligo $d(T)_{15}$ primer and AMV reverse transcriptase, second-strand synthesis was carried out with DNA polymerase I, after nicking the RNA strand with RNase H. 10μCi of $\alpha^{32}P$ dCTP was included in the reaction mix and samples were counted to monitor the efficiency of synthesis. This cDNA was stored briefly at -20°C before random prime labelling.

Phage lambda pod cDNA library construction

cDNA was synthesised from 2μg of poly(A) enriched RNA from 5 daf PP pods, using a Boehringer kit as described above, except that T_4 DNA polymerase was used (as per instructions) to blunt the ends of the cDNA after synthesis. This material was purified by extracting once with phenol/chloroform/isoamyl alcohol (25:24:1) and twice with chloroform/isoamyl alcohol (24:1), vortexing and spinning the tube at 12,000g for 3min at each step. The final aqueous phase was loaded onto a 5ml Sephadex G-50 column, equilibrated with TE + 0.1M NaCl. The excluded peak (judged with Gieger counter) was pooled, samples taken for scintillation counting, and the remainder ethanol precipitated.

Adaptors from blunt end to *Eco* RI containing a *Not* I site were ligated to the cDNA using 0.05units in a total volume of 10μl, overnight at 15°C. This was heated at 65°C for 10min and phosphorylated by incubating with 10μl 10mM ATP and 10units T_4 polynucleotide kinase, at 37°C for 30min. This was made up to 200μl with TE buffer and extracted with phenol/chloroform/isoamyl alcohol and chloroform/isoamyl alcohol, as above. The final aqueous phase was run through a 5ml Sepharose CL-4B column, in TE buffer + 0.1M NaCl and the excluded (radiolabelled) peak

pooled, ethanol precipitated, washed, dried and resuspended in 20μl H_2O.

This cDNA was cloned into λgt10 using the Promega Protoclone and Packagene systems. Following the supplier's instructions, test ligations were prepared using varying ratios of cDNA to *Eco* RI restricted λgt10 vector. Aliquots were then packaged with the supplied extract, dilutions of this used to infect an overnight culture of C600*Hfl* bacteria and plated. A scaled up ligation and packaging was then performed and titrated, yeilding a library of $6x10^5$ plaques. The whole library was amplified on plates, eluted with phage buffer and stored at 4°C.

Titration and screening of lambda phage libraries

Stored λ phage libraries were titrated prior to screening. 1μl of phage suspension was diluted with phage buffer and 100μl of these dilutions adsorbed, for 30min at 37°C, to an equal volume of the host bacteria (which had been grown overnight in LB + 0.2% maltose, 10mM $MgSO_4$ and antibiotic, if required, then stored at 4°C). Cells were then mixed with 3ml TB top agar at 45°C, poured over LB agar plates, incubated overnight and plaques counted.

Using the results from the titrations, large (20x20cm) plates were prepared by adsorbing the required quantity of phage and plating with 50ml TB over 300ml LB. Plates were incubated for 8-9hrs until discrete plaques could be seen but before confluent lysis, then cooled to 4°C. Duplicate filters were taken, overlaying the plates with nitrocellulose and, using a hole punch and ink, marking their position. These filters were processed, hybridised with ^{32}P labelled probe and autoradiographed. X-ray films were aligned with the plates, agar plugs containing areas of duplicating hybridisation removed with a cork borer and stored with phage buffer at 4°C.

Purification of phage lambda transformants

Phage suspended in the buffer surrounding the plugs removed after screening libraries was titrated as above. Plates with discrete plaques were screened as above, using 82mm diameter nitrocellulose filters. Duplicate hybridising regions were again removed and these titrated and screened until well separated single plaques could be removed and all plaques titrated from these were hybridised to by the probe.

Minipreparation of lambda phage DNA

Lambda phage was prepared using a Quiagen>lambda<kit, following the manufacturer's instructions. Phage was freshly titrated from stored stocks and eluted from a single plaque-plug with 100μl phage buffer. 500μl of an overnight host bacterial culture (which had been grown overnight in LB + 0.2% maltose, 10mM $MgSO_4$ and antibiotic if required) was added and incubated at 37°C for 20mins with shaking. This was innoculated into prewarmed 50ml LB + 10mM $MgSO_4$ and incubated for 7hr at 37°C with shaking.

If lysis had not occurred, 250μl of chloroform was added and incubation continued for a further 15min. Bacterial debris was precipitated by centrifugation at 8,000g for 10min and the supernatent removed and stored overnight at 4°C. The Quiagen "midi" method was then followed using their tip 100 columns to purify the phage DNA, finally resuspending in 100μl TE buffer. 5μl aliquots were used for restriction analysis.

General Isolation of pea genomic DNA

DNA to be used for restriction analysis and PCR amplification, rather than to make genomic libraries was prepared from frozen pea leaves and stipules using the rapid method of Ellis *et al.* (1984).

Genomic DNA digestion and electrophoresis for blotting

10μg samples of plant genomic DNA were restricted using at least 5units μg^{-1} DNA of restriction enzyme and incubating for 6hrs at 37°C, with shaking (around 100rpm). These were run overnight on 0.6% agarose gels, as normal.

Isolation of pea genomic DNA for use in genomic libraries

To recover genomic DNA of sufficient size suitable for partial digestion with *Sau* 3A, generating fragments of approximately 9-23kbp in length with *Sau* sites at both ends, the method of Graham (1978) was used. Leaves and stipules were taken from PP plants prior to flowering, frozen in liquid N_2 and stored at -80°C.

DNA extracted by this method was suspended in 50mM Tris-HCl 10mM EDTA pH 8.0 and incubated with 0.5mg ml^{-1} pronase P (self digested for 2hr) for 3hr at 37°C. The DNA was then purified by centrifugation, twice through caesium chloride gradients made up at 0.94g ml^{-1} CsCl with 0.1mg ml^{-1} ethidium bromide, centrifuging at 50,000g for 36hr. DNA bands, visualised under UV light (300nm), were removed, after puncturing the tube wall, through a 19 guage needle. DNA was extracted with amyl alcohol until colourless and dialysed against 50mM Tris 10mM EDTA pH8.0 with many changes. Finally, the solution was ethanol precipitated, washed, dried and resuspended in the same buffer as used for dialysis.

Size fractionation of genomic DNA

DNA was size fractionated using *Sau* 3A to yeild molecules in the range 9-23kbp suitable for cloning into the vector used. 180μg of genomic DNA was diluted to 1ml with Boehringer restriction enzyme buffer "A" to a final 1x concentration, by gentle inversion at 4°C for 2hr. To 166μl (30μg) aliquots of this were added various fresh dilutions of *Sau*

3A (0.4-2.5units) in 1-1.5μl volumes and these gently mixed and incubated at 37°C for 30min. The enzyme was denatured by heating at 70°C for 10min and the DNA phenol and chloroform/isoamyl alcohol extracted, ethanol precipitated, dried and resuspended at 0.5μg μl^{-1} in H_2O. 1μl aliquots were then electrophoresed through 0.4% agarose gel with lambda size markers.

Cloning of genomic DNA into lambda vector

Size fractionated genomic DNA was cloned into the Promega λGEM-12 *Xho* I half site arms vector and packaged using Promega Packagene extract following the supplied protocols. The arms of this vector have been cut with *Xho* I, and the ends partially filled in using dTTP and dCTP to leave TC- 5' sticky ends. *Sau* 3A digested genomic DNA is then partially filled in using dATP and dGTP to leave GA- 5' sticky ends (ie. -AG 3' sticky ends). These two species are now self-incompatible but can be ligated to each other, therefore avoiding both religation of arms to stuffer fragmant and multiple inserts of genomic DNA being cloned.

Sau 3A restricted DNA fractions, whose size when viewed on agarose gel was estimated to be around 25kbp, had their *Sau* 3A ends half filled in. 20μl of size fractionated DNA (10μg) was mixed with 20μl H_2O, 5μl 10x buffer containing dGTP and dATP (supplied with λ arms) and 5μl (5units) Klenow fragment of DNA polymerase I. This was incubated for 30min at 37°C and phenol and chloroform/isoamyl alcohol extracted. DNA was ethanol precipitated using an equal volume of 5M ammonium acetate and two volumes of -20°C ethanol. The tube was placed at -70°C for 30min then centrifuged at 12,000g for 15mins. The pellet was then washed and dried as usual and resuspended in 30μl H_2O. The DNA concentration of an aliquot was determined spectrophotometrically.

Test ligation mixes were then prepared using 0.5μg vector and various quantities of genomic DNA (62.5-750ng). The 5μl ligation mixes were incubated overnight at 4°C. 1μl aliquots from these were then packaged with 10μl of supplied extract for 2hr at 22°C, this was then diluted to 100μl and aliquots titrated and plated as usual using KW251 cells.

Multiple (5x) ligations were then set up using the optimum insert to vector ratio (62.5ng insert to 0.5μg vector). These were incubated as above, all of each ligation packaged with 50μl extract, pooled and an aliquot titrated. The remaining packaged extract containing $9x10^5$pfu was plated onto four 20x20cm plates and grown overnight. The phage was extracted with 100ml phage buffer per plate, overnight at 4°C. This solution was removed from the plates which were then rinsed with phage buffer, all this then centrifuged at 5000g for 10min and the supernat nts removed, mixed and stored at 4°C.

PCR amplification of lambda inserts

Insert from λgt10 was amplified by PCR to facilitate cloning. Oligonucleotide primers were synthesised complementary to sequences flanking the cloning site of this vector. A plug containing a freshly grown λ plaque was extracted with 1ml phage buffer and 20μl from this solution w boiled for 5min.

A 100μl reaction was set up using 0.1nM of each oligonucleotide, 5units of *Taq* polymerase, 10μl of the suppliers 10x buffer, 16μl of a 1.25mM (of each) dNTP mix and the boiled phage template. Mineral oil was used to overlay the solution and the tube placed in the incubation block. The sample underwent 29 cycles consisting of the following steps; denaturing - 90sec at 94°C, annealing - 1min at 50°C and extension - 2min at 72°C. This was followed by a single cycle of 90sec at 94°C, 1min

at 50°C and 5min at 72°C.

A 20μl aliquot of the reaction mix was electrophoresed through agarose gel, the band cut out and DNA isolated by electroelution. The amplified product was restricted using a large excess of *Eco* RI (50units), phenol and chloroform/isoamyl alcohol extracted, ethanol precipitated and subcloned into plasmid and M13 vectors.

PCR amplification of Birte genomic DNA

Oligonucleotide primers were synthesised against *legK* sequence (see fig. 21) and these were used to amplify sequence from 1.5μg of Birte genomic DNA. Reactions were conducted as above with the addition of 1μl of Stratagene "perfect match enhancer" per reaction. Amplified product was purified as above followed by treatment with T4 polymerase to blunt the ends (Maniatis *et al.*,1982).

S1 Nuclease mapping of transcription start points

The transcription start point of genes was determined by protecting single stranded DNA fragments extending into the 5' flanking sequence of these genes, by hybridisation with poly(A) enriched RNA from the tissue of interest. Non-hybridised regions were digested with S1 nuclease and the remaining fragments sized on sequencing gels.

The method was an adaptation of that of Favaloro *et al.* (1980). 10^5-10^6 cpm of 5'-end labelled DNA were mixed with 5μg of Poly(A) enriched RNA and 20μg of yeast tRNA (RNase free) and ethanol precipitated. The nucleic acid was resuspended in 10μl of hybridisation buffer (40mM PIPES pH6.4, 1mM EDTA, 0.4M NaCl, 80% formamide), placed at 75°C for 15mins and then transferred to a water bath at 50°C which was allowed to cool to 42°C and maintained at that temperature overnight. 300μl of ice cold digestion buffer (280mM NaCl, 50mM Na Acetate pH4.6,

4.5mM $ZnCl_2$, 20μg ml^{-1} phenol extracted herring sperm DNA, 200u ml^{-1} S1 nuclease) was added to the nucleic acid and this incubated at 37°C for 30min.

The reaction was stopped by the addition of 75μl of 2.5M ammonium acetate, 50mM EDTA, then extracted with an equal volume of phenol:choroform:IAA 25:24:1 and the aqueous phase precipitated with an equal volume of isopropanol. The nucleic acid was dried and resuspended in 5μl H_2O. An equal volume of sequencing dyes were added and the sample run on a manual sequencing gel. Standard M13 DNA was run as a size ladder, (a control experiment using a sequence primed at the same point as the 3' end of the DNA fragment protected, gave identical size to that estimated using M13 sequence), control experiments using no RNA were also performed with each probe.

Extraction of pod protein

Frozen pods were ground with a cooled mortar and pestle, thawed and extracted overnight at 4°C with gentle inversion at 100mg ml^{-1} in PBS containing 0.4mM EDTA, 2μg ml^{-1} leupeptin, 0.4mg ml^{-1} PMSF. Debris was precipitated by centrifugation at 10,000g for 10min, the supernatent removed and ammonium sulphate added to 100% saturation. After overnight incubation at 4°C with gentle inversion, protein was recovered by centrifugation at 17,000g for 30min. The pellet was resuspended in PBS, dialysed exhaustively against H_2O and lyophilised. Samples were resuspended at 5mg ml^{-1} in SDS sample buffer and boiled for 5min before polyacrylamide gel electrophoresis.

PBS-Phosphate buffered saline	150mM NaCl 9mM sodium phosphate pH 7.2
SDS sample buffer	0.2M Tris-HCl pH 6.8 2% SDS 10% Sucrose 20μg ml^{-1} Bromophenol Blue

Extraction of seed protein

Testas and embryonic axes were removed from mature seeds and the cotyledons ground in a mortar and pestle. Protein was extracted overnight at 4°C with gentle inversion, at 40mg ml^{-1} in SDS sample buffer (see above). Samples were boiled for 5min and debris was precipitated by centrifugation at 12,000g for 5min before loading the supernatent onto polyacrylamide gel.

Polyacrylamide gel electrophoresis

Proteins were electrophoresed on polyacrylamide gels, run in the dissociating SDS-PAGE buffer system (Laemmli, 1970). 8x10cm gels were cast and run in an ATTO AE-6450 apparatus, they were prepared as described by Hames (1981). Samples prepared as above and Mr standards were loaded into wells, β-mercaptoethanol added if required for reducing (about 1μl per well) and electrophoresed at 50V. Samples to be blotted were run in duplicate and gels cut in half after electrophoresis. Proteins were stained with Kenacid blue overnight and destained as required. Gels were photographed and dried between cellophane under vacuum.

Destain	50% Methanol 7% Acetic Acid
Stain	Destain + 0.5g l^{-1} Kenacid Blue

Western blotting of polyacrylamide protein gels

Protein was transferred from polyacrylamide gel to nitrocellulose filter using an ATTO AE-6670 semi dry western blotting apparatus. Onto the anode of the apparatus were placed in order: 2 sheets of 3MM paper (cut slightly bigger than the gel) soaked in 0.3M Tris-HCl pH10.4, 20% methanol, 0.1% SDS; 1 sheet of 3MM soaked in 25mM Tris-HCl pH10.4, 20% methanol, 0.1% SDS; 1 sheet of nitrocellulose soaked in H_2O; the gel; 1

sheet of 3MM soaked in 25mM Tris-HCl 40mM 6-amino-n-hexanoic acid pH 9.6, 20% methanol, 0.1% SDS; 1 sheet of cellophane soaked in H_2O and 2 sheets of 3MM soaked in 25mM Tris-HCl 40mM 6-amino-n-hexanoic acid pH 9.6, 20% methanol, 0.1% SDS. The cathode was placed on this and protein electrophoresed at 2.5mA cm^{-2} for 1hr. The gel was then stained with Kenacid blue to ensure all material had been transferred and the filter was blotted dry and stored.

Immunological detection of protein on Western blots

To detect proteins on nitrocellulose filters a Bio-Rad "Immun-Blot" kit was used, following the instructions provided.

The filter was incubated at room temp with the following solutions, the volumes given were for a (whole) 10x8cm gel: TBS, 20ml, 5min; 3% gelatin in TBS, 10ml, 1hr; 0.05% Tween-20 in TBS, 20ml, 10min; 0.05% Tween-20 1% gelatin in TBS + primary antibody (2.5μl anti-pectinesterase antiserum or 5μl anti-vicilin antibody), 10ml, overnight; 0.05% Tween-20 in TBS 20ml, 10min, twice; 0.05% Tween-20, 1% gelatin in TBS + 3μl alkaline phosphatase coupled anti-rabbit antibodies, 10ml, 2hr; 0.05% Tween-20 in TBS, 20ml, 10min, twice; TBS, 20ml, 10min; Colour developement solution, 10ml, until stained (usually 30min); H_2O, 5min, twice. Filters were blotted dry and stored in the dark.

TBS-Tris buffered saline 20mM Tris-HCl pH7.5
0.5M NaCl

Extraction and ammonium sulphate precipitation of root protein

Roots were removed from pea plants and rinsed with H_2O containing β-mercaptoethanol (0.5ml l^{-1}), then frozen in liquid N_2 and lyophilised. The dried material was ground in a mortar and pestle and extracted with

25ml g^{-1} tissue in PBS containing 432mg l^{-1} PMSF, 10mg l^{-1} leupeptin, 20mM β- mercaptoethanol, first homogenising with a polytron for 20sec at maximum speed then shaking at 4°C for 4hr. Debris was removed by centrifugation at 25,000g. for 20min. Ammonium sulphate precipitation was performed by adding the desired amount of ammonium sulphate slowly to stirring solution at 4°C. The solution was left stirring at this temperature for at least 90min and the precipitate was collected by centrifugation as above.

Cyanogen bromide cleavage of root protein

Trial CNBR cleavage was performed on purified major root protein by the method of Allen (1981). Protein was suspended at 5mg ml^{-1} in 70% formic acid and an equal weight (to protein) of CNBr (suspended in acetonitrile at 2g ml^{-1}) added. The solutions were mixed, flushed with N_2 and incubated at room temperature in the dark for 24hr. The reaction was then diluted at least 15x with H_2O and lyophilysed.

Reduction and carboxymethylation of root protein

To break disulphide bonds and prevent their reformation, root protein was reduced and carboxymethylated by the method of Konigsberge (1972). Protein was suspended at 2% in 6M Guanidine-HCl, 0.5M Tris-HCl pH8.2, then flushed with N_2 and heated at 50°C for 30min. A 50x molar excess of DTT (to protein) was added, the solution again flushed with N_2 and incubated for 4hr at 50°C. After cooling to room temperature a 1.5x molar excess over DTT of iodoacetamide was added and this incubated at room temperature for 40min. The resulting solution was then dialysed twice against a large volume of 0.1M ammonium bicarbonate and lyophilysed.

CHAPTER FOUR: CONVICILIN RESULTS

Restriction mapping and sequencing of the genomic clone

The genomic clone λJC4 and its subclone pJC4-100 had previously been isolated and identified as containing a convicilin related sequence (Ellis *et al.*,1986); this convicilin gene was designated *cvcA*. Restrictions were performed on pJC4-100 and the resulting restriction map is presented in fig. 4. Restriction fragments which contained regions hybridising to the convicilin cDNA (pCD59; Casey *et al.*,1984) were subcloned into plasmid vectors and subsequently into M13. These M13 subclones were sequenced manually, and a map of the sequencing strategy is also shown on fig. 4. The genomic subclone pJC4-100 is approximately 14kbp in length with 8.5kbp lying 5'and 3kbp 3' of the convicilin gene.

The sequenced region (fig. 5) is 3284bp in length, 589bp are 5' of the start codon and 437bp 3' from the stop codon. The predicted aa sequence is also presented on fig. 5. It was deduced from homology to vicilin (Lycett,*et al.*,1983a), and the presence of an ORF at the 5'end. There are five introns in *cvcA*, their positions were established by comparison of the genomic sequence with that of the convicilin cDNA and the vicilin aa sequence. The introns are 151,103,103,88 and 97bp in length and intersperse exons of 661,176,75,324,283 and 194bp respectively. This results in a coding sequence of 1713bp encoding 571 aa.

The *cvcA* encoded protein

There is a leader sequence at the N-terminus of the *cvcA* encoded protein, which has a hydrophobic core LLLFL(G)IIFLA and consensus patterns (von Heijne,1983) suggest cleavage after the 28th residue. The mature protein would then be 543aa in length and have an N-terminal sequence NYDEGSETRV-, which is identical to that determined from the

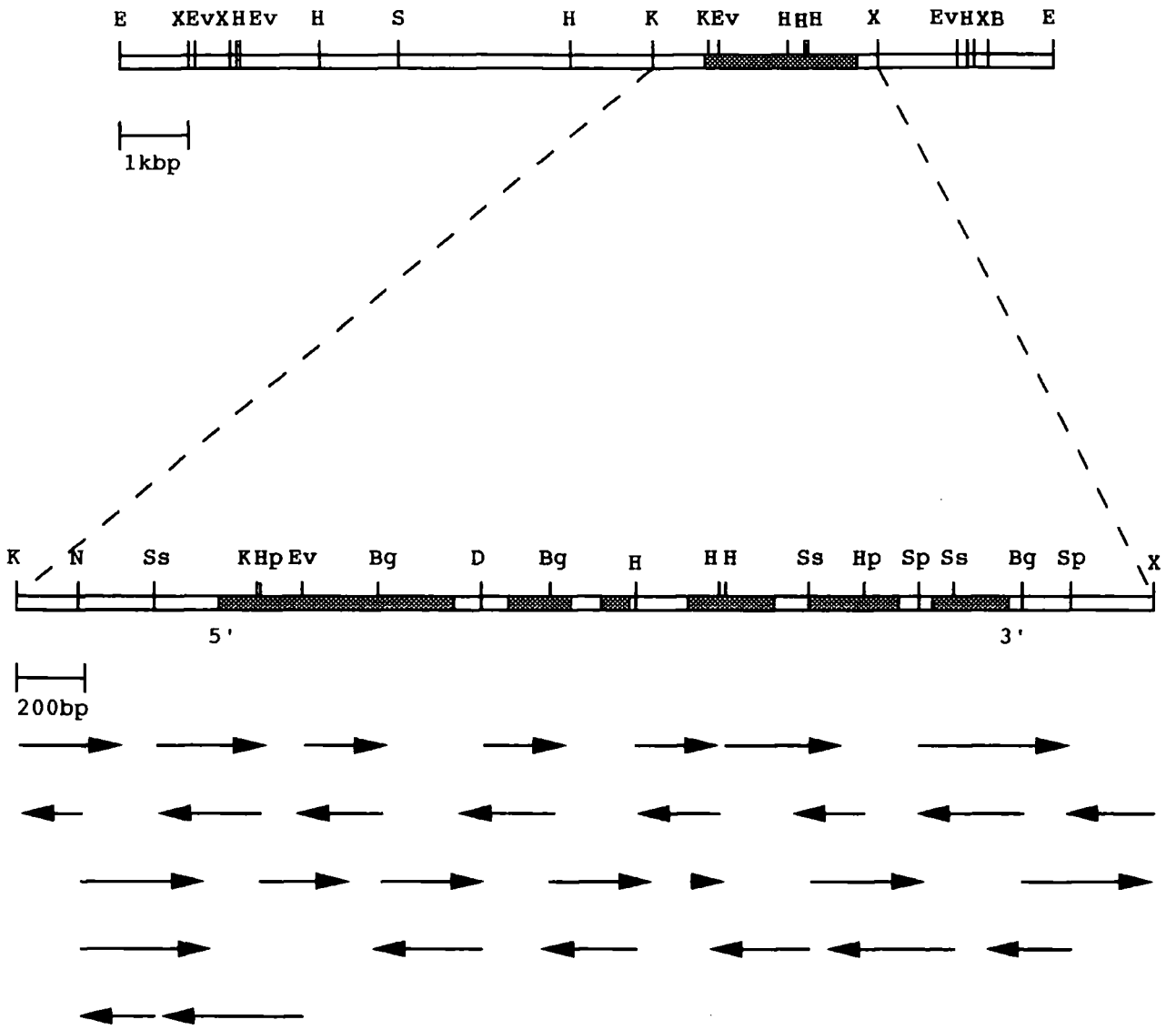


Fig. 4 Restriction map of pJC4-100 genomic subclone (above) and the sequenced region (below). The *cvcA* coding region is highlighted. Restriction enzymes are abbreviated as: B, *Bam* HI; Bg, *Bgl* II; D, *Dra* I; E, *Eco* RI; Ev, *Eco* RV; H, *Hind* III; Hp, *Hpa* II; K, *Kpn* I; N, *Nsi* I; S, *Sal* I; Sp, *Ssp* I; Ss, *Sst* I; X, *Xba* I. For clarity not all *Dra* I and *Ssp* I sites are shown on the restriction map of the sequenced region. Arrows represent individual sequencing runs.

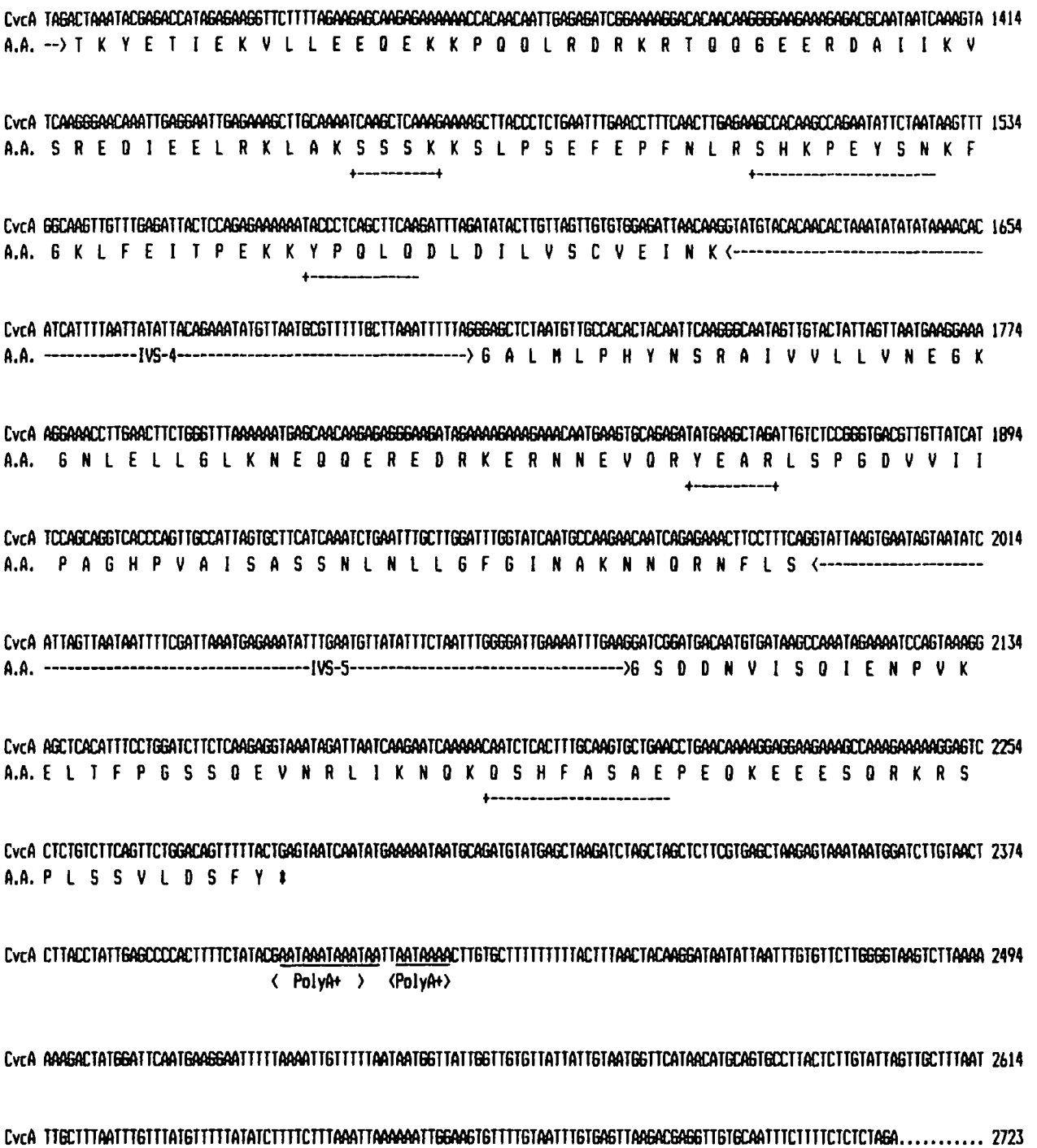


Fig. 5 Nucleotide and derived amino acid sequences of *cvcA*. The signal peptide cleavage site is marked (:). Peptides sequenced from the protein (Bown *et al.*,1988) are marked, with residues differing from those predicted shown. Polyadenylation signals and regions conserved between the 5' flanking regions of vicilin family genes are underlined. Restriction sites used to generate the "convicilin specific" probe are also marked. The sequence against which the *cvcA* specific oligonucleotide, used to probe N blots, was made, together with the corresponding sequence from pCD59 (shown above) is boxed.

protein (Bown *et al.*,1988).

Fig. 6 Amino acid composition of the mature protein predicted by *cvcA* and that determined from convicilin (Croy *et al.*, 1980a)

	Residues predicted		Mole % predicted	Mole % determined
alanine	18		3.3	4.2
cysteine	1		0.2	0.1
aspartate	23)			
		) 59	10.9	11.6
asparagine	36)			
glutamate	80)			
		) 113	20.8	22.1
glutamine	33)			
phenylalanine	20		3.7	3.3
glycine	27		5.0	5.9
histidine	12		2.2	2.2
isoleucine	24		4.4	3.9
lysine	43		7.9	8.2
leucine	49		9.0	8.7
methionine	1		0.2	0.1
proline	25		4.6	5.5
arginine	53		9.8	8.2
serine	40		7.4	6.4
threonine	13		2.4	2.6
valine	27		5.0	4.5
tryptophan	3		0.6	ND
tyrosine	15		2.8	2.6

ND = not determined

The amino acid composition predicted by this sequence (fig. 6) is very similar to that determined on the purified convicilin protein (Croy *et al.*,1980a) and results in a predicted Mr of 63,986 for the mature protein encoded by *cvcA*.

Expression of convicilin in the developing cotyledon

In order to study the pattern of expression of convicilin in the developing seed, a northern blot with total RNA from pea cotyledons at different developmental stages was probed, washing to 0.1xSSC, 0.1%SDS at 50°C. The 700bp *Sst* I to *Bgl* II fragment of *cvcA* which covers the 5' region of the gene, encoding aa sequence not found in vicilin (fig. 5),

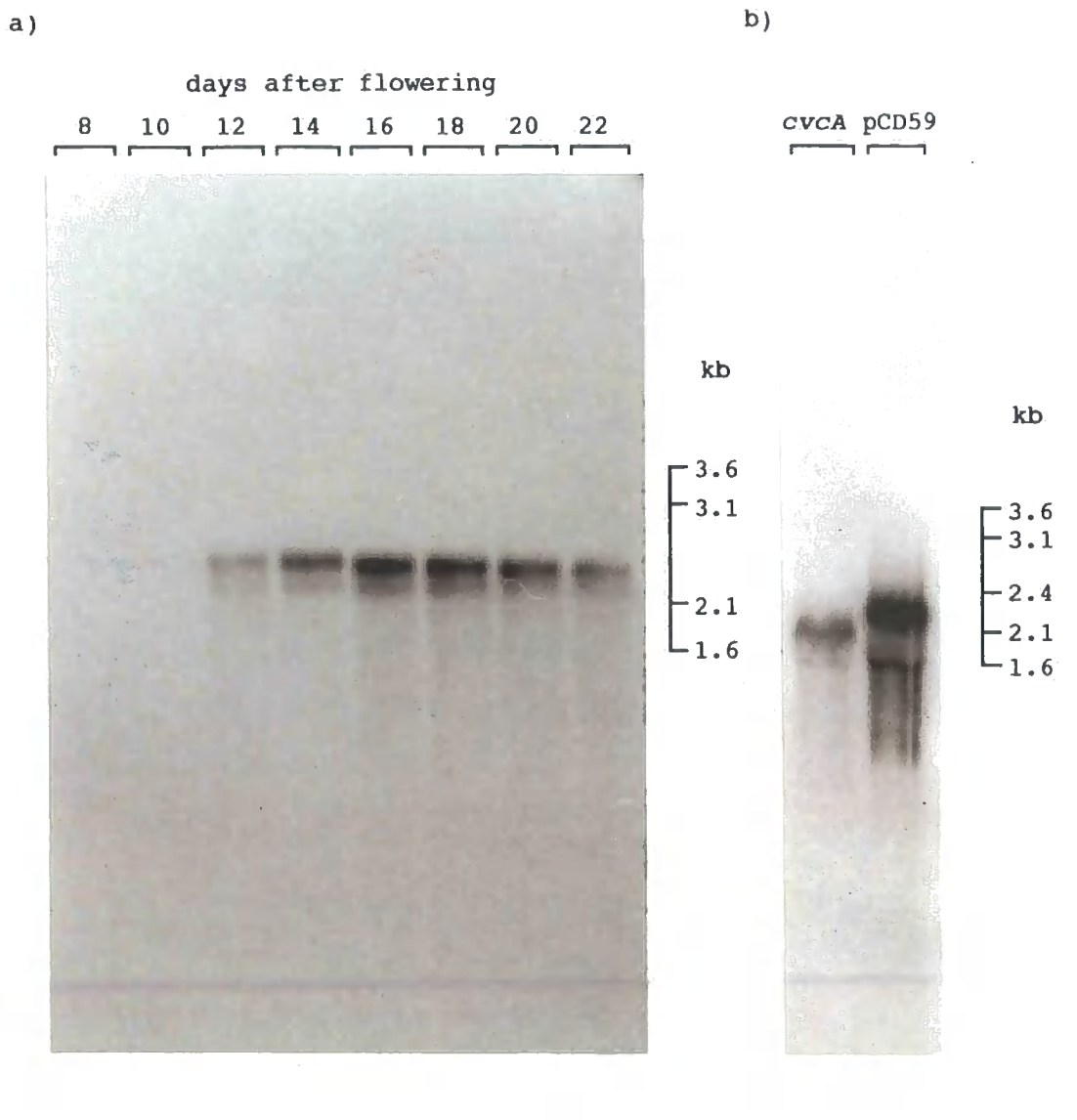


Fig. 7 Northern blots probed with convicilin; a) cotyledon RNA through development probed with the "convicilin specific" *Sst* I - *Bgl* II fragment from *cvcA*. b) 18 daf cotyledon RNA probed with oligos specific to *cvcA* or pCD59 - convicilin cDNA. The position standard RNA ran to on the original gels is marked.

was used as a "convicilin specific" probe. The result (fig 7a), is that two mRNA species are detected, at 2.7 and 2.5kb. The relative intensity of these two bands remains constant through development, scanning the film with a laser densitometer confirms this and gives a ratio of 3:1 for the 2.7 and 2.5 kb bands respectively. mRNA could not be detected by the convicilin probe in RNA from 8 daf. coyledons, but it could be seen at 10 daf. and increased to a peak at 16-18 daf. then declined to 50% of peak level at 22 daf.

To account for the two mRNA species, oligonucleotides were synthesised which were specific for either the cDNA or the genomic sequence (fig. 5). These oligonucleotides were end-labelled with ^{32}P and used separately to probe tracks from the same blot containing RNA from 18 daf. cotyledons. Prehybridisation and hybridisation were performed at 25°C and the filters were washed to 5xSSC at 37°C. The cDNA specific oligo hybridised to a mRNA species 2.3kb in length and the *cvcA* specific oligo to an mRNA 2.1kb in length. Both oligos also detected a faint band at 1.8kb (fig. 7b).

Probing genomic blots for convicilin sequences

Genomic blots were performed to investigate the convicilin gene subfamily and to ascertain whether there were differences between the pea line from which the genomic clone (but not the cDNA) was isolated, Dark Skinned Perfection (DSP) and the standard line used in this laboratory, Feltham First (FF), from which the RNA used above had been extracted. Digests of DNA from both lines were probed with the same convicilin specific probe used for the northern blots and the filter was washed to 0.1xSSC at 65°C.

Both pea lines gave the same result (fig. 8): there are two genes present, which lie on restriction fragments of identical size in the two

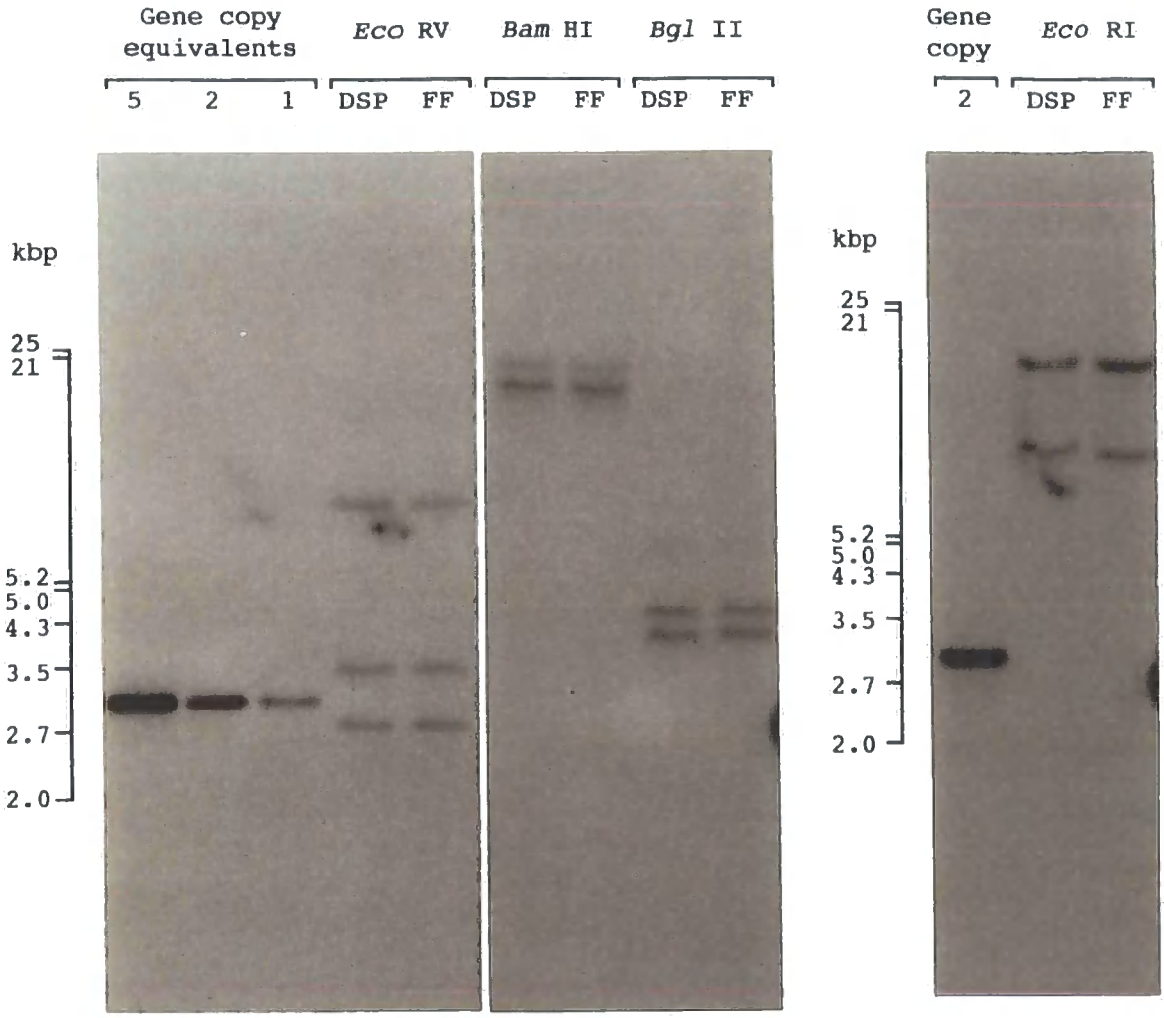


Fig. 8 Blots of pea genomic DNA from FF and DSP lines restricted with various enzymes probed with a convicilin specific fragment of *cvcA*. Gene equivalent amounts of *cvcA* were also run. The position of size marker DNA bands on the original gels is indicated.

genomes. Restriction fragment sizes in the *Eco* RI and *Eco* RV digests, equivalent to one gene copy, correspond to that predicted by the genomic clone (fig. 4): *Eco* RI, 13kbp; and *Eco* RV, 7.1 and 3.6 kbp. The correspondence in sizes for those digests where the size of fragments in the genomic clone is known demonstrates that no rearrangements have occurred during the cloning procedures.

S1 Nuclease mapping of the transcription start of *cvcA*

To confirm that this gene is expressed and to locate the transcription start point, S1 nuclease mapping experiments were conducted. The 700bp *Asp* 718 and the 660bp *Nsi* I - *Eco* RV fragments

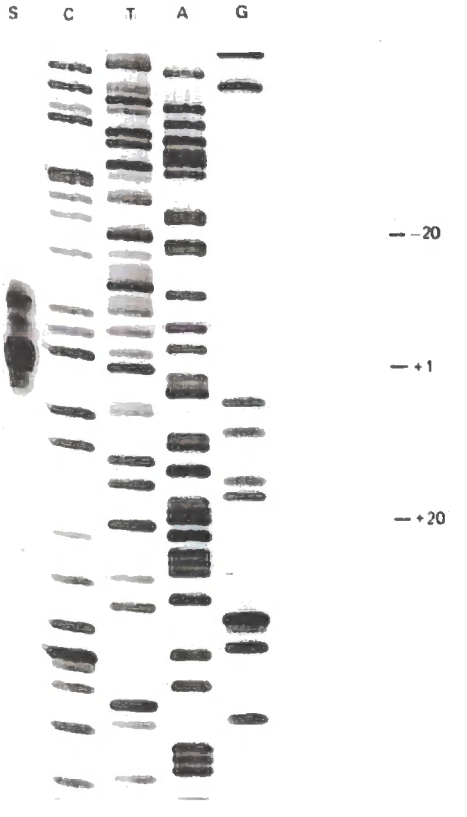


Fig. 9 S1 nuclease mapping of the transcription start point of *cvcA*. The protected *Asp* 718 fragment is run in track S and the other tracks contain the corresponding region of the DNA sequence. The base corresponding to the strongest protected fragment is labelled +1.

covering the 5' flanking region of *cvcA* and 5' coding sequence (fig. 5) were end labelled for use in these experiments (*Asp* 718 is an isoschizomer of *Kpn* I which generates 5'-protruding ends for labelling). When denatured and hybridised to bulked (mid development, 14-15daf) cotyledon poly(A) enriched RNA, fragments of 119 to 130b from the *Asp* site and 254 to 266b from the *Eco* RV site were protected by the RNA (figs. 9 & 16). The strongest protected bands were 121b in length from the *Asp* site and 256b from the *Eco* RV site. These distances were used to assign the transcription start point. The sequence of *cvcA* (fig. 5) is numbered from this point which is the underlined base in the sequence TTCATCCATCTAAAG. This is 28bp 5' from the translation start point and 38bp distal to the TATA box.

Expression of *cvcA* in transgenic tobacco seeds

In order to identify the product of *cvcA* and to compare it with the product of a second convicilin gene whose sequence proved to be homologous to the cDNAs (Newbigin *et al.*,1990), *cvcA* was used to transform tobacco. The 3.1 kbp *Nsi* I to *Xba* I fragment containing *cvcA* coding sequence with 0.4kbp 5'- and 0.4kbp 3'- flanking sequence (fig. 4) was subcloned into pDUB126A. This vector is a derivative of pBR322 carrying gentamycin/kanamycin resistance and a multipurpose cloning site (Edwards, 1988).

This vector was mobilised into *Agrobacterium tumifaciens* and these used to transform tobacco, *Nicotiana tabacum* cv. Petit Havana Str-r_1 (SR1), leaf discs (Jobes,D. and Croy,RRD., unpublished work). Plants were regenerated on selective media, grown to maturity and the seeds collected (Jobes,D. and Croy,RRD., unpublished work). Convicilin expression was assayed by ELISA (Bryden,J. and Gatehouse,JA., unpublished work).

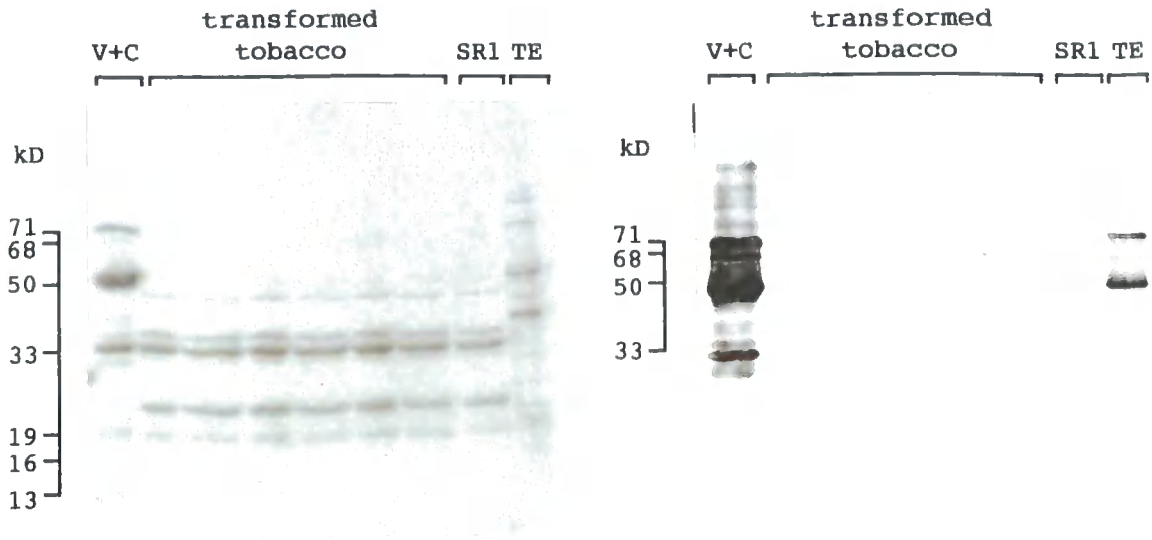


Fig. 10a Gel (left) and anti-vicilin probed Western blot (right) of duplicate gel containing seed extracts from tobacco plants transformed with *cvcA*, non-transformed tobacco (SR1) and pea (TE). A sample of vicilin and convicilin (V+C) was also run and the position and Mr of polypeptides in this marked.

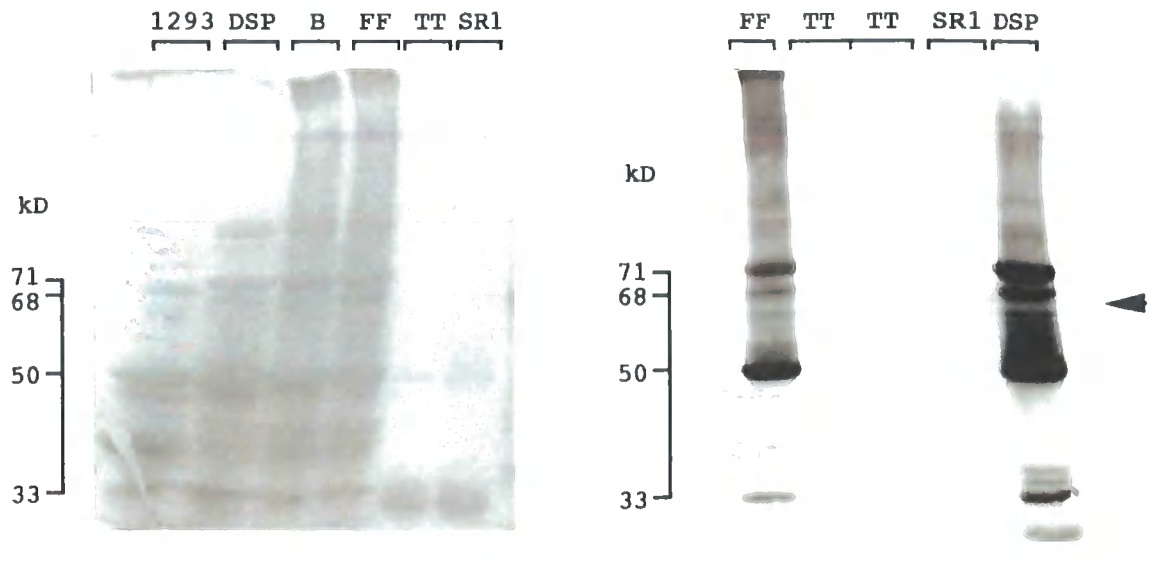


Fig. 10b Gel (left) and anti-vicilin probed Western blotted portion of the same gel (right). The gel contained seed extracts from tobacco plants transformed with *cvcA* (TT), non-transformed tobacco (SR1) and pea lines 1293, DSP, Birte (B) and FF. The position and Mr of polypeptides in the pea seed extract is indicated. The arrow marks the position of a contaminant detected by the anti-vicilin antibodies across the entire filter.

5 seeds from plants with detectable levels of convicilin or non-transformed plants were homogenised in SDS sample buffer and extracted protein electrophoresed through polyacrylamide gels. Western blots of these gels were then probed with anti-vicilin antibodies which had previously been shown to cross-react with convicilin (Croy *et al.*,1980a). Protein extracted from transformed plants contained a band detected by these antibodies of approximately 65k Mr which was not present in the untransformed tobacco (fig. 10). This band is equivalent to a polypeptide in the pea vicilin and convicilin fraction and cotyledon total extracts, detected by these antibodies, and below the main (71k Mr) convicilin band. Bands of around 50k Mr were also present in the transformed but not the untransformed tobacco extracts and a band of around 47k Mr was present in all tobacco extracts.

CHAPTER FIVE: CONVICILIN DISCUSSION

General remarks

Since the majority of the investigations involving *cvcA* were completed and published (Bown *et al.*,1988, q.v., bound with this work), the nature of the products of *cvcA* and the gene encoding the cDNA used to isolate it has been clarified by the work of Newbigin *et al.* (1990) and further work by this author. The results will therefore be discussed in the light of these later observations.

The *cvcA* gene encodes a minor, lower Mr species of convicilin

The results of the western blots of seed protein from transgenic tobacco plants transformed with the complete coding sequence of *cvcA* and 0.4kbp of 5' flanking sequence, demonstrate that the inserted gene encodes a vicilin related protein species with a Mr of around 65,000. The bands generated are of identical mobility to a protein detected by the vicilin antibobies in a vicilin + convicilin extract from pea seeds and in pea seed total extract (fig. 10).

The protein encoded by *cvcA* is therefore of lower Mr than the major convicilin protein at 71k (Croy *et al.*,1980a) encoded by the gene transformed into tobacco by Newbigin *et al.*(1990). The gene (p5.1) used by these workers predicts a 578 residue product with a resulting Mr of 68.2k. Its sequence is homologous to that of *cvcA* except for two repeats in the inserted sequence (see below) of p5.1, which account for the size difference of the predicted products. The sequence of this (p5.1) gene is, however, 98% identical (Newbigin *et al.*,1990) to that of the cDNA pCD59 (Casey et al.,1984), which is identical to that of the cDNA pCD75 used to isolate the genomic clone on which *cvcA* lies (Ellis *et al.*,1986, Domoney & Casey,1990).

It appears then that the p5.1 gene encodes the larger convicilin

species estimated experimentally to be 71k Mr (but predicted by the sequence to be 68.2k) and that the *cvcA* gene product is that estimated to be 65k and predicted by *cvcA* to be 63.9k. The discrepancies between the predicted Mr and that observed on SDS-PAGE are presumably due to the highly charged nature of the residues in the inserted sequences (see below) affecting the mobilities, a factor more pronounced in the higher Mr form, with its longer inserted sequence (167 vs 121aa).

A cDNA, pCD72, from a distinct class of vicilin related genes whose translation product occurs at a similar Mr to that of *cvcA* has recently been reported (Domoney & Casey, 1990). This cDNA encodes a vicilin like sequence with internal methionine residues but insufficient data is available to determine whether its high Mr (compared to vicilin) is due to a similar inserted sequence to that of the convicilin genes. Although the gene product is of a similar size to that encoded by *cvcA* it has been shown that a methionine rich vicilin precursor of this Mr is post-translationally cleaved (Chrispeels *et al.*,1982). It seems likely that the polypeptides encoded by the pCD72 type gene are processed to lower Mr species and that they do not encode the 65k Mr species ascribed here to be the product of *cvcA*.

The tobacco seed protein extracts on western blots from plants transformed with *cvcA* (fig. 10) also contain bands at around 50k Mr not present in untransformed plant extract. These bands are presumably the result of processing of the convicilin by a mechanism not found in pea (Chrispeels *et al.*,1982) but which also affected the product of the other convicilin gene, p5.1, in transgenic tobacco (Newbigin *et al.*,1990). The possibility that the 65k Mr band is the result of cleavage, by these tobacco specific mechanisms, of the larger Mr convicilin, is argued against by both the difference in size of the proteins predicted by the two genes and the detection of an identically

sized band, by vicilin antibodies, in pea seed total extract.

A 47k Mr protein was detected by the vicilin antibodies in all tobacco extracts. This presumably represents a vicilin-like storage protein from tobacco seeds and has been detected in non-transformed tobacco plants by other workers (Higgins *et al.*,1988, Newbigin *et al.*,1990).

The convicilin gene subfamily in pea

The genomic blot probed with the 5' "convicilin specific" region of *cvcA* (fig. 8) indicates the presence of two closely related genes in the FF and DSP genomes. One gene is represented by the fragments predicted by *cvcA* and the other lies on an approximately 8kbp *Eco* RI restriction fragment which was hybridised to by the cDNA, pCD59, in DSP DNA (Domoney & Casey, 1985, Ellis *et al.*,1986). These two gene copies are genetically linked as they have been found to segregate together in crosses (Ellis *et al.*,1986). The (p5.1) gene also lies on a similarly sized (7.5kbp) *Eco* RI fragment in pea line "Greenfeast" genomic DNA (Newbigin *et al.*,1990).

Other workers have noted the presence, in their genomic digests, of fragments hybridised to weakly by convicilin probes. In *Eco* RI digests bands at 7kbp (Domoney & Casey, 1985) or 5.5kbp (Newbigin *et al.*,1990) have been reported but no such bands were visible using the *cvcA* probe.

Expression of the *cvcA* gene

S1 mapping experiments indicate that *cvcA* is expressed in pea seeds. The results of the northern blotting work reflects the finding that *cvcA* represents the smaller Mr species of convicilin. The pattern of expression of both species is identical. The result confirms those obtained previously on northern and dot blots (Chandler *et al.*,1984,

Boulter *et al.*,1987), that peak expression of convicilin mRNA occurs later in seed development than that of vicilin and more closely resembles that of legumin.

The pattern and ratio of intensity of the hybridisation through development is also mirror ed at the protein level with the 65k Mr species consistantly less plentiful than the 70k species and both accumulating later in development than vicilin (Gatehouse *et al.*,1982b). The ratio of the amounts of protein from the two convicilin species in the mature FF seed is approximately five to one when the upper and lower band (respectively) are compared (5.3 +/- 0.7 by scanning by laser densitometer, kenacid blue stained gels with varying quantities of protein loaded - data not shown). The ratio of these species is similar in many pea lines, including DSP (from which λJC4 was isolated), however, Birte (from which pCD59 & 70 were isolated) does not appear to contain a polypeptide of this Mr (fig. 10).

Convicilin gene *cvcA* coding sequence

The coding sequence of this gene predicts a precursor polypeptide of 571aa residues in length. The sequence predicts the removal of a signal peptide which is supported by the observation that convicilin undergoes post-translational modification (Higgins & Spencer, 1981). It is likely that a signal peptide is removed from the precursor as is the case with the other major storage proteins of pea, vicilin (Lycett *et al.*,1983a) and legumin (Lycett *et al.*,1884a).

The aa composition of the mature protein predicted by this sequence matches very well that determined previously on the convicilin protein (fig. 6). The low figure predicted for E+Q presumably results from the lack of the glutamate-rich repeats in *cvcA* compared with the gene encoding the larger convicilin species (see below). Apart from this the

figures reflect the similarity of the two species. This includes two sulphur containing residues, C and M, neither of which are found (outside the leader sequence) in vicilin (Lycett *et al.*,1983a). The predicted sequence is also confirmed by its agreement with the sequence of tryptic peptides isolated from the protein (Bown *et al.*,1988).

Comparison with other convicilin encoding DNA sequences

The cDNA clone pCD59 has been identified as coding for convicilin (Domoney & Casey,1983) and is identical, over the region of overlap (Domoney & Casey, 1990), with pCD75, the cDNA used to select the *cvcA* genomic clone (Ellis *et al.*,1986). However, when the *cvcA* and pCD59 sequences are compared, they are not the same. Over the 591bp of the cDNA sequence (Casey *et al.*,1984) there is 94% identity to *cvcA*, there are 18 silent base changes and 16aa changes. Two deletions occur in the cDNA sequence with respect to the genomic clone, one of 18bp (6aa) and another of a single codon. The larger deletion occurs in a region previously noted for its variability within the vicilin gene family, around the α:β processing site of the 47k Mr protein (Lycett *et al.*,1983a) (fig. 11).

The partial genomic sequence reported by Newbigin *et al.* (1990) is more homologous to the pCD59 sequence, with only four base differences in the 321bp overlap. The partial cDNA (pPS15-28) sequence which extends the genomic sequence in a 3' direction, however, despite having only 4 nucleotide differences with the genomic clone in the 516bp overlap, is different from both *cvcA* and pCD59. This difference occurs within the same region as the difference between *cvcA* and pCD59 (fig. 11).

vicB	E H E K E T Q H R R S L K:D K R Q Q S Q E E N V I V K L
cvcA	E Q E K K P Q Q L R D R K R T Q Q G E E R D A I I K V S
pCD	E Q E K D R K R R Q Q G E E T D A I V K V S
pPS	E Q E K E P Q Q R R A I V K V S

Fig 11 Comparison of the predicted amino acid sequences from: Vicilin gene *vicB* (Boulter *et al.*,1990), (: denotes the position of the potential α:β cleavage site in vicilin); convicilin gene *cvcA*; convicilin cDNA pCD59 (pCD) (Casey *et al.*,1984); and convicilin cDNA pPS15-28 (pPS) (Newbigin *et al.*,1990). The sequences start at residue 13 of exon four of the genomic sequences and finish at the end of pCD59

This result suggests there may be more than two species of convicilin-like gene present in the pea genome. Although only two were detected on the genomic blot (fig. 8), other workers have reported the presence of at least one other weakly hybridised band (see above). An added complication is the differring pea lines used to isolate the various cDNAs, pCD59 and 75 from Birte, pSP15-28 from Greenfeast, and genomic clones *cvcA* (λJC4) from DSP, and p5.1 from Greenfeast (Domoney & Casey, 1983, Chandler *et al.*,1984, Domoney *et al.*,1986a, Newbigin *et al.*,1990).

Apart from the above mentioned region, the pPS15-28 cDNA and p5.1 genomic sequences are otherwise homologous to *cvcA* (95% identity when gaps are introduced for homology - Newbigin *et al.*,1990). The other main difference being the occurrence of two repeats at the 5' end of the inserted sequence (see below), one of 96bp and the other 75bp.

Comparison with other legume storage protein genes:

Inserted sequence

On comparison of the convicilin coding sequence with that of vicilin, both at the aa and nucleotide level, (fig. 12 shows a dot matrix comparison of predicted aa sequences from *cvcA* and vicilin 50k Mr cDNA) it is immediately obvious that, although homologous to vicilin

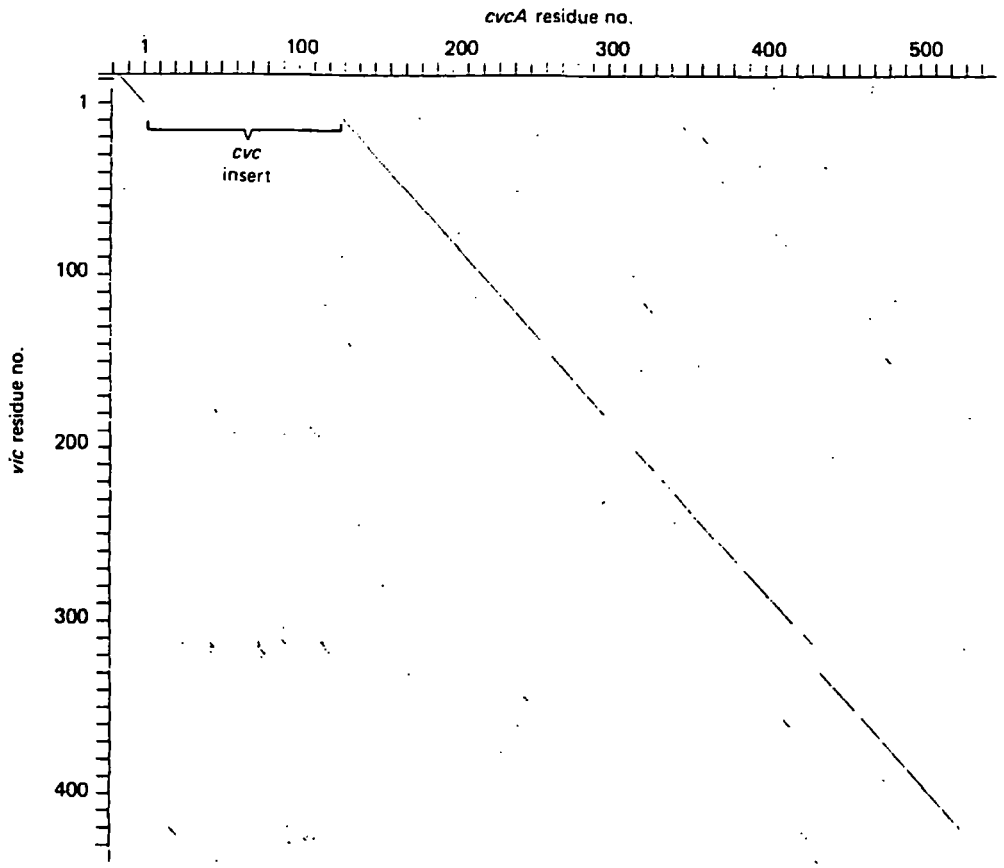


Fig. 12 Dot matrix comparison of the amino acid sequence encoded by *cvcA* with that of vicilin. (Gatehouse, *et al.*,1984, Boulter *et al.*,1990) Using the matrix of Staden (1982), sequences were compared over a span of 8 residues with a minimum score of 102

over an extensive (500aa) region, the *cvcA* gene contains an insertion near the 5' end of the coding sequence. The exact position of the end points of this insertion are difficult to determine due to the poor homology to vicilin in this region. It appears the insertion occurs between the third and sixth aa of mature vicilin and is thus represented by residues 4 to 124 (nucleotides 122-484 on fig. 5) of the mature convicilin.

This inserted sequence has little homology to the remainder of the convicilin and vicilin coding sequences, it consists of a large proportion of hydrophilic residues, with a total of 80 charged aa (out of 120). At the nucleotide level it is A+G rich (78% A+G). These features are reminiscent of the C-terminal region of the α-subunit of

legumin (Lycett *et al.*,1984b, Gatehouse *et al.*,1988, Rerie *et al.*,1990) but when compared at both the nucleotide and aa levels no direct sequence homology exists. Variable repeats are found in this region of the α-subunit of the pea legumin gene family. Although the inserted region of *cvcA* lacks any direct repeats, the inserted region of the other convicilin gene (p5.1, Newbigin *et al.*,1990) does contain two repeated sequences (see above).

Inspection of the sequence flanking the *cvcA* inserted sequence and dot matrix comparison failed to detect any inverted repeats in this region. These findings would argue against this insertion being a result of the action of a transposable element as these features are generally associated with such an event (Freeling,1984).

The α'-subunit of β-conglycinin (soybean convicilin) also contains an inserted sequence near its N-terminus when compared to other vicilin type genes (Doyle *et al.*,1986). The conglycinin insertion is 174aa in length when compared to phaseolin (*Phaseolus* vicilin) and occurs between residues 6 and 7 of the mature phaseolin polypeptide. Both at the nucleotide and aa levels these insertions are similar, that in conglycinin also being A+G rich and encoding hydrophilic and charged aa. Despite this, as with legumin, when the sequences are compared, no significant sequence homology can be found. This, together with the observation that the remainder of the *cvcA* coding sequence is closer to that of vicilin than it is to conglycinin (see below), implies that two separate insertion events have occurred.

A cotton vicilin also contains an insertion, at a similar position to *cvcA*, with a high proportion of glutamate, glutamine and arginine (93 out of 162 residues), but also with a high cysteine content (12 residues in 162) arranged in C-XXX-C peptides (Chlan *et al.*,1986). Insertion and rearrangement (through duplications) of such hydrophilic sequences

appears therefore to be a frequent method of mutation of plant storage protein genes. Work on soybean β-conglycinin has shown that mutation of the hydrophilic N-terminal region can be tolerated to a much greater extent than mutation of the conserved C-terminal region, during oligomer assembly *in vitro* (Lelievre *et al.*, 1992).

The coding sequence 5' to the insertion in *cvcA* is homologous to that of vicilin, the predicted aa leader sequence is a much better match to that of vicilin than to legumin or other storage protein genes. This confirms that *cvcA* does in fact result from an insertion into the vicilin sequence, rather than a fusion of the 3' region of a vicilin gene with the 5' region of another, unrelated, gene.

The vicilin-like sequence

The region of *cvcA* 3' of the inserted sequence has close homology to vicilin and the only regions where homology does break down are around the potential α:β subunit processing site of vicilin (residues 297/298 bases 1563/4 on fig. 5) and the β:γ processing site where convicilin has a 6aa deletion with respect to vicilin 47k Mr (in the region of residues 425-431 bases 1835-1850). These regions have previously been noted for their variability in the vicilin gene family and between related gene families (Lycett *et al.*,1983a, Casey *et al.*,1984). This sequence homology between vicilin and *cvcA* would account for the similar properties shared by these proteins (Croy *et al.*,1980a).

When compared to other vicilin related storage protein genes, phaseolin (Slightom *et al.*,1983) and conglycinin (Doyle *et al.*,1986), the non-inserted coding sequence shows a similar pattern of homology as it does to vicilin, with the same regions of variability around the potential processing sites. However, when the nucleotide sequences were compared, exon by exon, using a computer alignment program, convicilin

and vicilin showed the greatest homology (73%), compared with convicilin to conglycinin (67%) and convicilin to phaseolin (59%). It is interesting to note that although convicilin shows greatest homology to vicilin this is lower than the homology between phaseolin and conglycinin (78%) and this confirms the position of convicilins as a subfamily of the vicilin family rather than being within the main vicilin family. These results also reflect the position of the genera *Glycine* and *Phaseolus* being within the same tribe (Phaseolae), whereas *Pisum* is separate from them in the Vicieae tribe (Polhill, 1981).

The intervening and 3' flanking sequence

Comparison of the sequences of *cvcA* and vicilin genes show that the intron positions are conserved and correspond to those of phaseolin and conglycinin. Neither the inserted sequence in convicilin nor that of conglycinin contain any introns. The intron boundaries follow the consensus pattern (Breathnach *et al.*,1978) and, as previously noted for plant genes (Slightom *et al.*,1983), they are A+T rich. When the introns of *cvcA* are compared to those of a vicilin gene, *vicB* (Boulter *et al.*,1990), intron by intron, the homology (43%) is much lower than that of the exons (73%).

The 3' non-coding sequence of *cvcA* extends for 3kbp on the genomic clone, the region beyond that sequenced will not be discussed further. Within the 437bp sequenced 3' to the stop codon there are a number of polyadenylation signals conforming to the consensus A/G ATA A_{1-3} (Messing *et al.*,1983) including the multiple overlapping type found in other plant genes including storage proteins (Lycett *et al.*,1983b). This region bears little sequence homology to the same region of the other convicilin gene beyond 45bp 3' from the stop codon.

The 5' flanking sequence of *cvcA*

The transcription start point of *cvcA* as determined by S1 nuclease mapping covers a region of nucleotides 24-35 5' of the start codon. Despite the fact that this region includes two sequences conforming to the core of the consensus CGCATCA for transcription start regions in dicot storage proteins (Joshi,1987), the S1 mapping experiments repeatedly gave the underlined base in the sequence CATCCATCT as the strongest transcription start point. This unusual result was also found when similar experiments were conducted the other convicilin gene (Newbigin *et al.*,1990). A TATA box is present 38bp 5' from the transcription start point, which is at the far end of the range of distances usually found between these features, 32+/-7 (Joshi,1987).

In an attempt to discern possible regulatory regions, the *cvcA* sequence was compared to that of other vicilin family genes, with the aid of a computer alignment program. A weakly conserved sequence occurs 25-50bp upstream of the TATA box, this includes the potential regulatory sequence CCAAAT which is conserved between all genes compared. This sequence has a widespread occurrence in this location especially amongst animal genes (Messing *et al*,1983) and has been shown to be required for maximum activity of the nopaline synthase gene in *Agrobacterium* infected tobacco tumor cells (Shaw *et al.*,1984), although other workers have not found this sequence to have a significant effect in other transcriptional assays *in vitro* (Grosveld *et al.*,1981) or in transformed plant tissue (Morrelli *et al.*,1985).

A conserved sequence starting 88bp 5' of the TATA box (fig. 5) occurs in *cvcA* and the other genes compared, and has been termed the "vicilin box" (Gatehouse *et al.*,1986). This sequence may be divided into two regions; 1) at the 5' end, a highly conserved C-rich sequence of 8bp, GCCACCTC, which is also found in pea and *Vicia faba* legumin genes

(Lycett *et al.*,1985, Baumlein *et al.*,1986) and 2) a less well conserved sequence of 34bp in which the base composition reflects that of the whole 5' flanking sequence. It has been suggested that this latter region, which is not found in the legumin genes, may play a role as a gene family and tissue specific transcriptional enhancer (Gatehouse *et al.*,1986). A sequence, A/T/C AACACA A/C A/T/C, similar to the latter portion of the "vicilin box" has also been identified in the 5' region of all soybean seed protein genes (Goldberg,1986) and it was suggested that this plays a role in regulating seed protein gene expression.

Further 5' from the vicilin box are two other sequences conserved between *cvcA*, phaseolin and conglycinin, at - 190 to -183 from the transcription start (CTCAACCC) and at -293 to -285 (GATCGCCGC). No sequences matching the purine-pyrimidine (RY) repeats with consensus CATGCATG that have previously been identified in most legume storage protein genes (Dickinson *et al.*,1988) could be found in either the 5' flanking sequence of *cvcA* or the other convicilin gene, p5.1. Comparing these two regions reveals close homology (72% identity) continuing to the end of the *cvcA* sequence (589bp from start codon). No obvious differences could be found to account for the different levels of mRNA and protein produced by the two genes.

Comparison of these sequences within the vicilin gene family should help identify tissue- or family- specific elements. However, functional testing in transgenic plants and DNA-protein binding assays would be required to assess the effect of elements found by sequence analysis. Northern blots through development, probed with *cvcA* (fig 7a) and vicilin cDNAs (Gatehouse *et al.*,1982b), show that the patterns of expression within the family are not identical and there could possibly be other temporal regulatory elements controlling this feature of expression.

G+C rich sequences such as those identified above, including that at the 5' end of the "vicilin box", have been found to be associated with many different promoters (Dynan & Tjian, 1985, Dierks *et al.*,1983) and have been identified as DNA-protein binding sites (Dynan & Tjian, 1985). *In vitro* transcription assays using deletion mutants have shown such regions to be necessary for maximum transcription (McKnight & Kingsbury, 1982). A similar region has also been identified in the conglycinin gene where four repeats of the sequence A A/G/C CCCA have been found in the -159 to -257 bp region 5' of the transcription start. Deletion of these causes a twenty fold decrease in the expression of this gene in transformed petunia plants (Chen *et al.*,1986) and the inclusion of these elements caused a tissue specific enhancement of reporter gene activity of up to twentyfive fold (Chen *et al.*,1988).

Work using DNA-protein binding assays on legume seed storage protein genes has so far failed to demonstrate binding of nuclear protein to elements identified by sequence comparison; the vicilin or legumin boxes (Bustos *et al.*,1989, Jordano *et al.*,1989, Meakin & Gatehouse, 1991) or the RY repeats (Riggs *et al.*, 1989). It has been suggested however, that less abundant, more specific proteins bind to these regions and that these are masked experimentally by the presence of the more abundant, general factors observed so far (Meakin & Gatehouse, 1991).

The factor which was found to bind to the 5' flanking region of *legA* (LABF1) binds to regions 5' of the legumin box which are required for activation of this gene in transgenic tobacco (Shirsat *et al.*,1990). Binding to LABF1 by these regions of *legA* is competed against by the *Kpn* I fragment from *cvcA* (fig. 4) which contains all the 5' flanking region sequenced and this result reinforces the general tissue-specific nature of LABF1. No highly conserved sequences could be found between *legA* and

cvcA (Meakin & Gatehouse, 1991). Continuing work should yield the specific sequence bound to by LABF1 in the near future and it will be interesting to compare this to the sequence of *cvcA* used as a competitor.

Sequencing of the vicilin 47k encoding cDNA, pLG1.63

The pea cotyledon cDNA pLG1.63 had been isolated by homology to a cDNA, pDUB7, encoding 47k Mr vicilin and restriction mapped (Gatehouse, 1985). Fragments from the insert were subcloned into M13 and subjected to manual DNA sequencing. The whole insert was sequenced in both directions and all restriction sites sequenced through (fig. 13). The sequence is presented in fig. 14; it is 1596bp in length and includes a 44 base poly(A) tail at the 3' end.

The predicted aa sequence (also on fig. 14) was deduced by homology to vicilin (Lycett *et al.*,1983a) and the presence of an ORF at the 5' end. The cDNA covers the entire coding sequence of 1314bp, encoding 438aa, it has 12bp 5' to the start codon and 226bp from the end of the coding sequence to the point of attachment of the poly(A) tail. The N-terminal sequence of the mature vicilin subunit encoded by this subfamily of genes has been determined (Lycett *et al.*,1983a). Removal of the signal peptide of 24 residues results in a polypeptide 414aa in length with a Mr of 47,163.

The pLG1.63 sequence contains 11 base differences with respect to the previous cDNA sequence (Lycett *et al.*,1983a) resulting in 9aa changes, none of these are within the potential processing regions or glycosylation site.

Isolation and sequencing of *vicJ*

The genomic clone, λJC3 had previously been isolated, restriction mapped and shown to contain sequences homologous to a cDNA, pCD4, encoding a 47k Mr species of vicilin (Domoney & Casey, 1983, Ellis *et al.*,1986). The gene on this clone has been designated *vicJ* to differentiate it from the genes encoding 50k Mr vicilin, *vic A,B,C* etc.

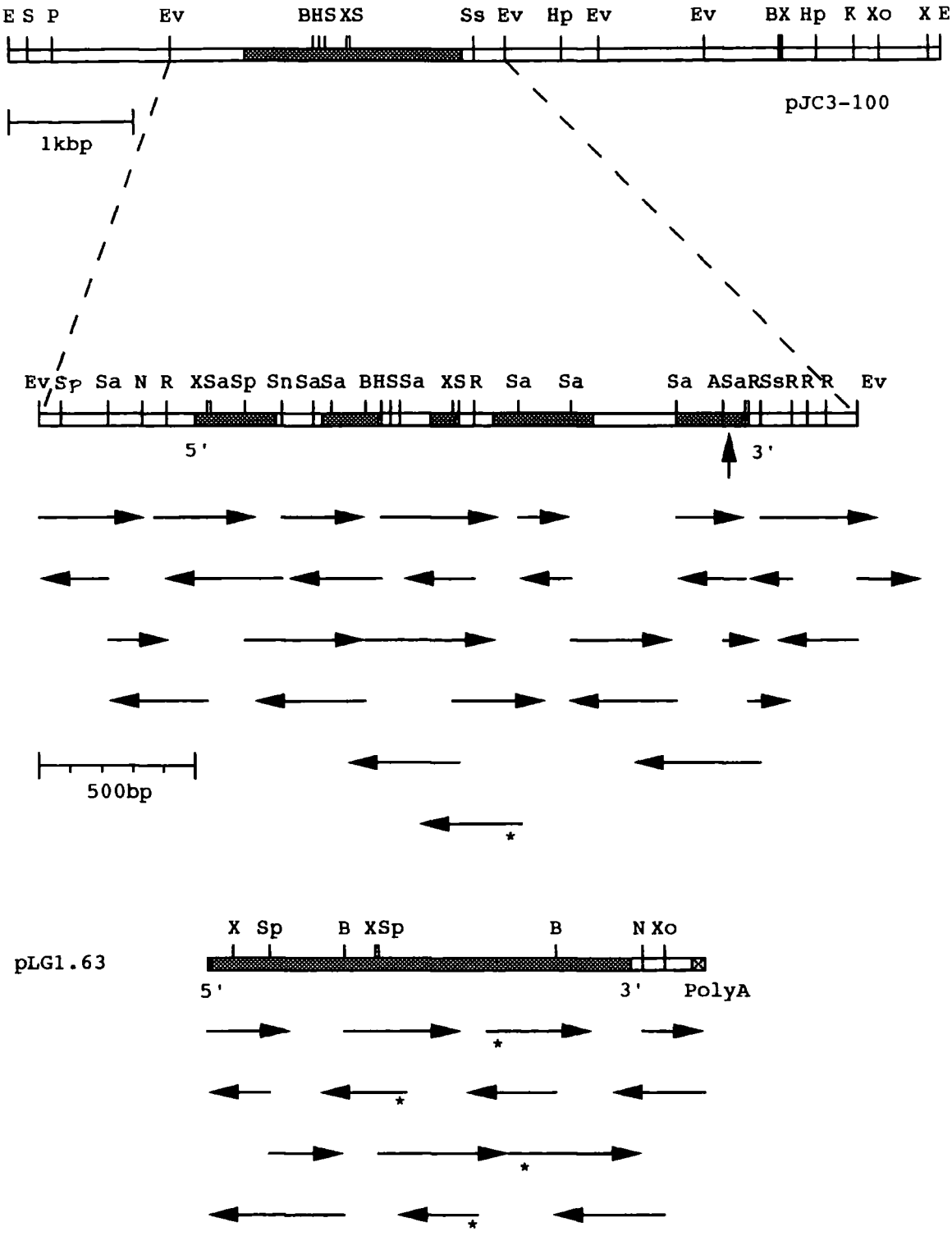


Fig. 13 Restriction maps of pJC3-100 genomic subclone (above), the sequenced region from it and pLG1.63 cDNA (below). Restriction enzymes are abbreviated as: A, *Alu* I; B, *Bgl* II; E, *Eco* RI; Ev, *Eco* RV; H, *Hind* III; Hp, *Hpa* I; N, *Nsi* I; P, *Pst* I; R, *Rsa* I; S, *Sph* I; Sa, *Sau* 3A; Sn, *Sna* BI; Sp, *Ssp* I; Ss, *Sst* I; X, *Xba* I; Xo, *Xho* I. For clarity not all *Ssp* I and *Alu* I sites are shown on the sequenced region of pJC3-100. The coding regions are highlighted, the vertical arrow shows the point of divergence of *vicJ* from the cDNA sequences. Horizontal arrows below the restriction maps represent individual sequencing runs, those marked * are primed from oligonucleotides synthesised complementary to the determined sequence.

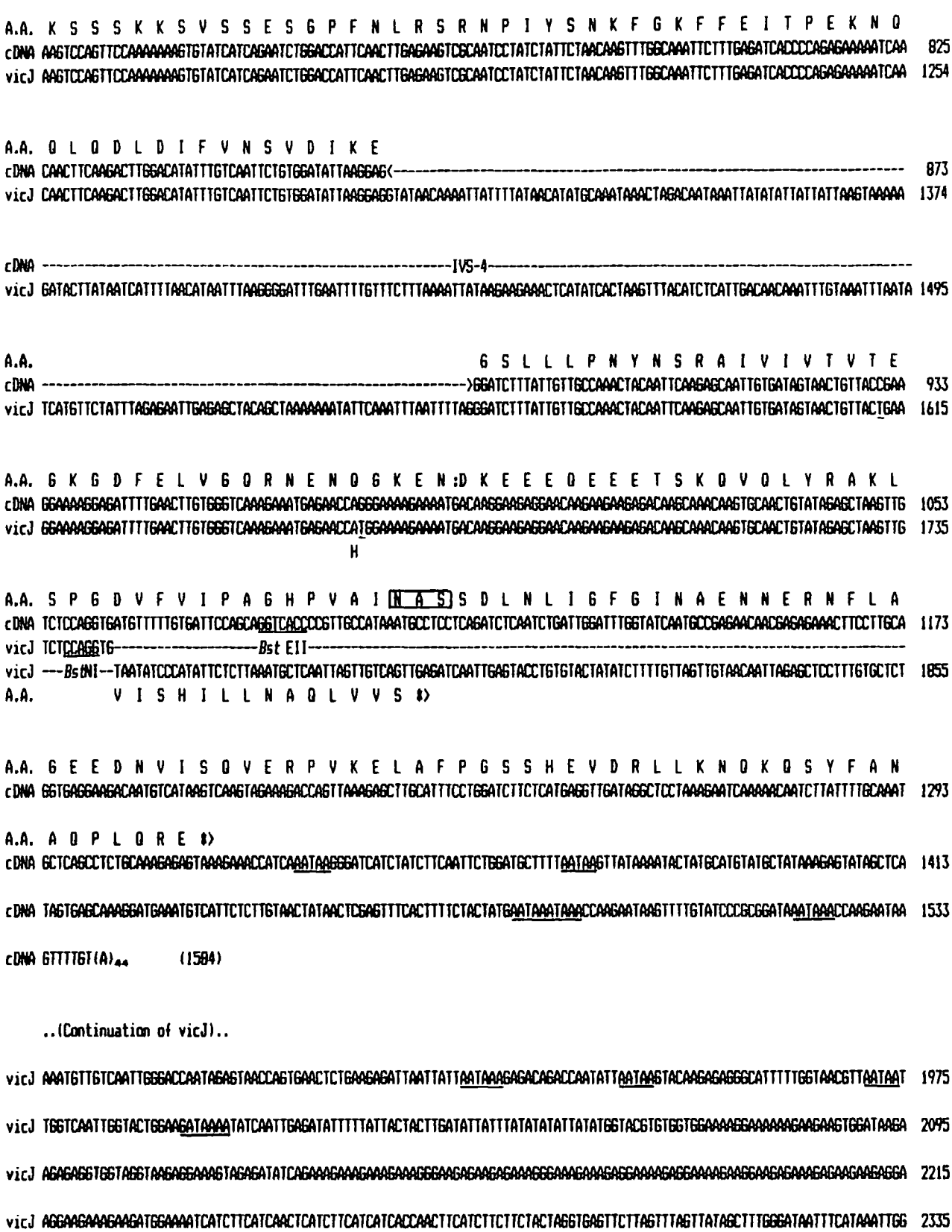


Fig. 14 Nucleotide sequences of *vicJ* and pLG1.63 (cDNA). Derived amino acid sequence from pLG1.63 is shown above the DNA sequence and residues deviating from this in *vicJ* are shown below. The signal peptide and subunit cleavage sites are marked (:). The potential N-glycosylation sequence and nucleotides protected during S1 mapping are boxed. Polyadenylation signals and regions conserved between the 5' flanking regions of vicilin genes are underlined. Restriction sites used to generate probes are also marked.

Restrictions were performed on pJC3-100, a subclone containing the 7.4kbp *Eco* RI fragment from λJC3 (Ellis *et al.*,1986), and from the sizes of bands on agarose gel, a restriction map was deduced (fig. 13).

The sequence homologous to the cDNA lies on a 2.7kbp Eco RV fragment, this fragment was subcloned into pUC8 vector and restriction mapped. Fragments from this were subcloned into M13 and subjected to manual DNA sequencing. A diagram of the clones sequenced is also presented on fig. 13 and the sequence obtained appears on fig. 14. The 2629bp *Eco* RV fragment was sequenced in both directions and all restriction sites sequenced through. An additional 208bp 3' from the *Eco* RV site are also presented, sequenced in the 5' to 3' direction (with respect to the gene) only.

The predicted aa sequence (also on fig. 14) was deduced by homology to pLG1.63 (see above). The nucleotide sequence of *vicJ* deviates from that of the pLG1.63 and other vicilin 47K cDNA clones, pDUB4 and pDUB7, at base 190 of exon 5 (see fig. 14). This diverged sequence contains an in-frame stop codon 42bp 3' from the point of deviation. The aa sequence predicted by this diverged sequence bears no homology to that of the vicilin protein or that predicted by vicilin cDNAs for this region. The DNA sequence extends for approximately 590bp 3' from the point of divergence from the cDNA and none of this sequence shows homology to the 3' end of the vicilin 47k cDNAs.

When the non-diverged (vicilin encoding) sequence of *vicJ* is compared with the cDNA sequence, it can be seen that there are four introns within the *vicJ* coding sequence, 148, 175, 106, and 252bp in length, these result in five exons of 294, 177, 81, 324, and 231bp. These intron positions are confirmed by their identical positions in *cvcA* (qv.) and *vicB* (Boulter *et al.*,1990). There are only 3 base differences in this (1075bp) region between *vicJ* and pLG1.63, resulting

in 2aa substitutions.

Experiment to locate the 3' coding sequence missing from *vicJ*

To ascertain whether sequence homologous to the cDNAs continuing the interupted vicilin coding sequence lies further 3' along the genomic clone, a southern blot containing digests of the genomic subclone, pJC3-100, the genomic clone, λJC3, and pLG1.63 was prepared. This was probed with the 450bp *Bst* EII to *Eco* RI fragment from pLG1.63 which extends over the area of vicilin sequence missing from the region of *vicJ* sequenced (fig. 14). The probe hybridised only to the original cDNA insert (result not shown) indicating that no homologous sequence lay on the 5.6kbp of the genomic clone 3' of the sequenced region.

The gene subfamily encoding 47k Mr vicilin

In order to investigate the nature of the vicilin 47k gene subfamily, genomic blots were performed using FF and DSP DNA, washing to 0.1xSSC at 65°C, the results are shown on fig 15. When the 5' portion of pLG1.63 up to the *Bst* EII site (encoding vicilin sequence up to the point of divergence with *vicJ*, fig. 14) was used as a probe (fig, 15a), bands were hybridised to indicating the presence of several homologous genes in the pea genome. Fragments hybridized to included those predicted by λJC3 (*Eco* RV - 2.7kbp, *Eco* RI - 7.4kbp, *Bam* HI - 12kbp, *Bgl* II - 3.8) demonstrating that no rearrangements had occurred in the cloning and subcloning of *vicJ*.

Genomic blots probed with the same region of *vicJ* (data not shown) and the 3'(*Bst* EII - *Eco* RI) end of pLG1.63 (fig. 15b) gave similar results for gene copy number. However, when probed with the 3' end of *vicJ* (the 390bp *Bst* NI to *Eco* RV fragment, fig. 14) genomic digests contained multiple bands hybridised to by this probe (fig. 15c). Many of

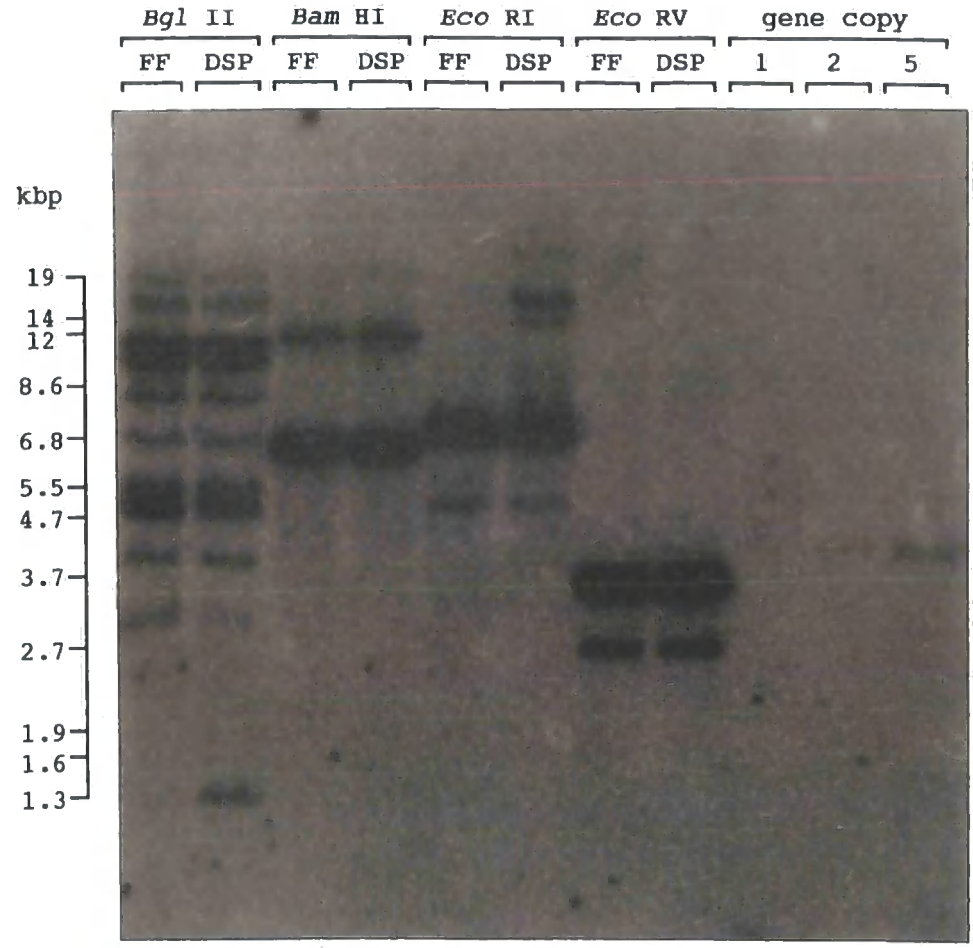


Fig. 15a Genomic digests of Feltham First (FF) and Dark skinned Perfection (DSP) DNA probed with the 5' (*Eco* RI-*Bst* EII) region of pLG1.63, 47k Mr vicilin encoding cDNA. Gene copy equivalent amounts of pLG1.63 were run. The position standard DNA migrated to on the original gel is indicated.

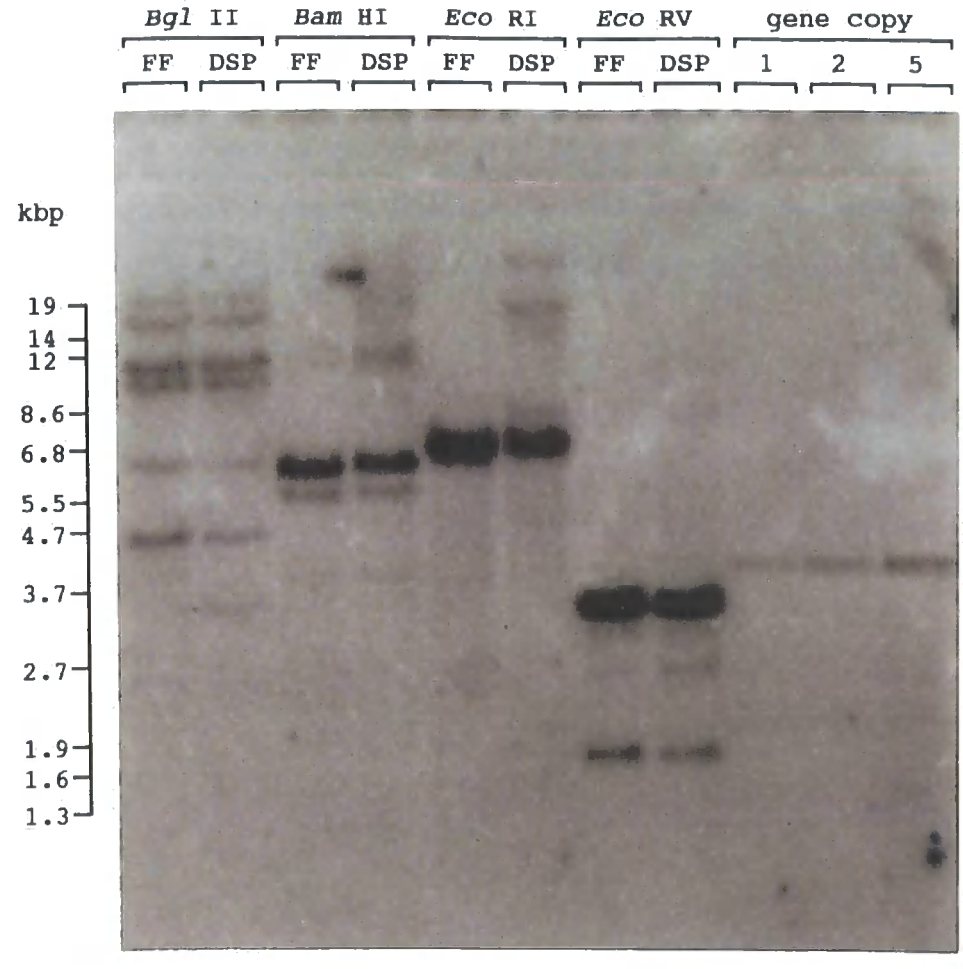


Fig. 15b Genomic digests of FF and DSP DNA probed with the 3' (*Bst* EII-*Eco* RI) region of pLG1.63, 47k Mr vicilin encoding cDNA. Gene copy equivalent amounts of pLG1.63 were run. The position standard DNA migrated to on the original gel is indicated.

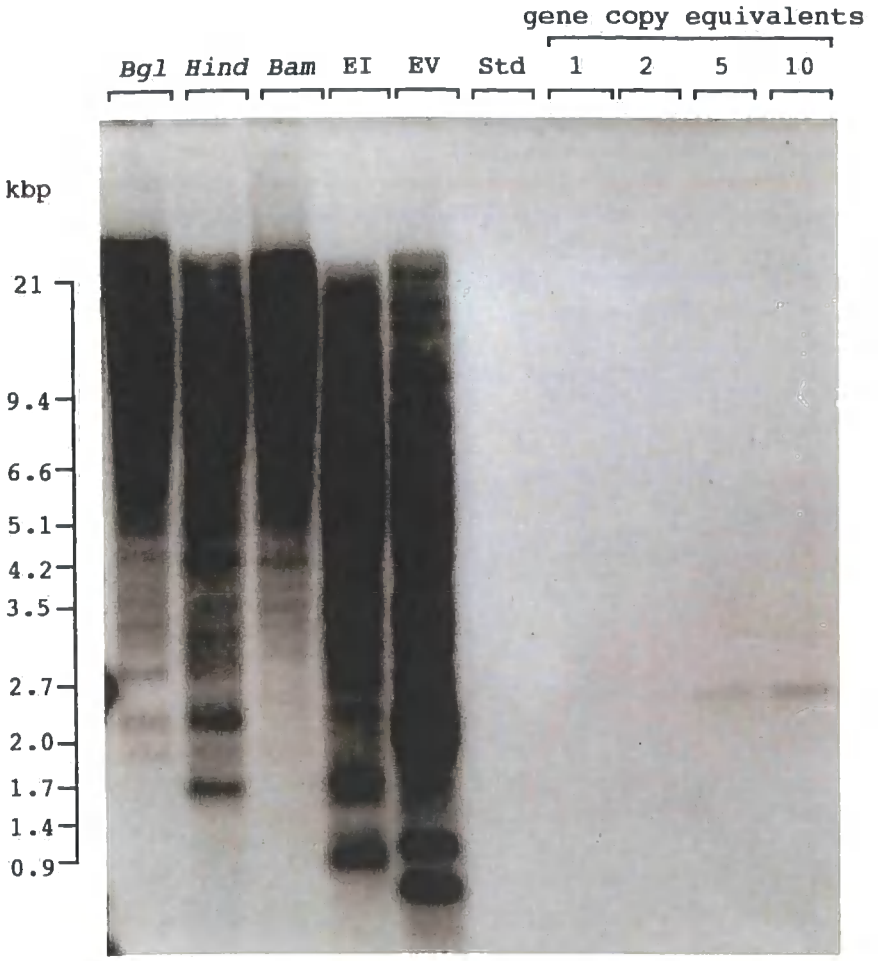


Fig. 15c Genomic digests of FF DNA probed with the 3' (*Bst* NI-*Eco* RV) diverged region of *vicJ*. Restriction enzymes used were *Bgl* II, *Bam* HI, *Hind* III, *Eco* RI (EI) and *Eco* RV (EV). Gene copy equivalent amounts of *vicJ* were run. The position standard DNA (Std) bands migrated to on the original gel is marked.

these bands had an intensity much greater than that of the 10 gene copy equivalent. When this probe was used on digests of the λJC3 and pJC3-100, only the fragments containing the probe fragment were hybridised to, indicating that the sequence of this region is not repeated on the genomic clone (data not shown).

Expression of the *vicJ* gene subfamily

An S1 mapping experiment was performed to determine the 5' end of the 47k vicilin encoding mRNAs. The 616bp *Ssp* fragment, covering the first 185bp of putative translated sequence and its adjacent flanking sequence (fig. 13), was hybridised to bulked (mid development 14-15daf) cotyledon poly(A) enriched RNA. The hybrids were subjected to S1 nuclease treatment and sized on a sequencing gel against DNA of known sequence. The mRNA protected fragments of 227-235b in length (fig. 16), placing the transcription start 23-31bp 3'of the TATA box, 41-50bp 5' of the start codon. Less intense fragments were protected in the size range 48-78bp. The fragment of *vicJ* used is identical in sequence to pLG1.63 in the region of their overlap (fig. 14). A result indicating an identical start point was also obtained using the 300bp *Sau* 3A fragment (fig. 13) from the same region of *vicJ* (data not shown).

N blots of total RNA from 18daf DSP and 14daf FF cotyledons probed with both the 5' and 3' regions of pLG1.63 contained only one mRNA species hybridised to by these probes (fig 17). This 2.0kbp species is identical in size in both pea lines and with both probes.

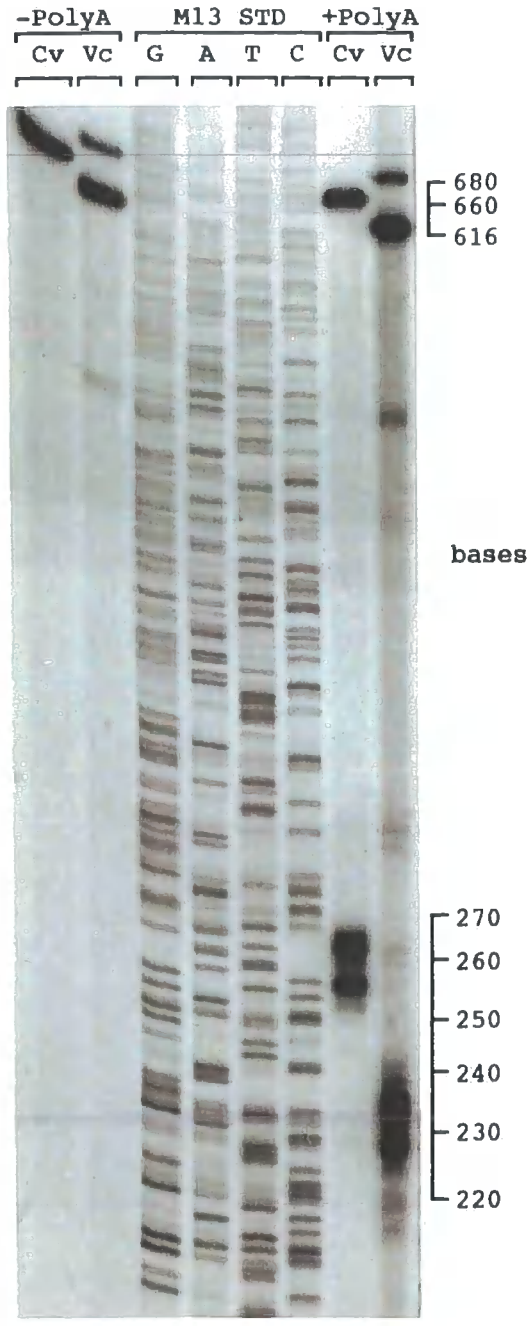


Fig. 16 S1 mapping experiment with *vicJ* (Vc) and *cvcA* (Cv). Labelled fragments from both genes were digested with S1 nuclease in the presence (+PolyA) or absence (-PolyA) of seed Poly(A) enriched RNA. Sizes of bands in the M13 standard (M13 STD) are indicated together with fragments representing the full-length (undigested) probes.

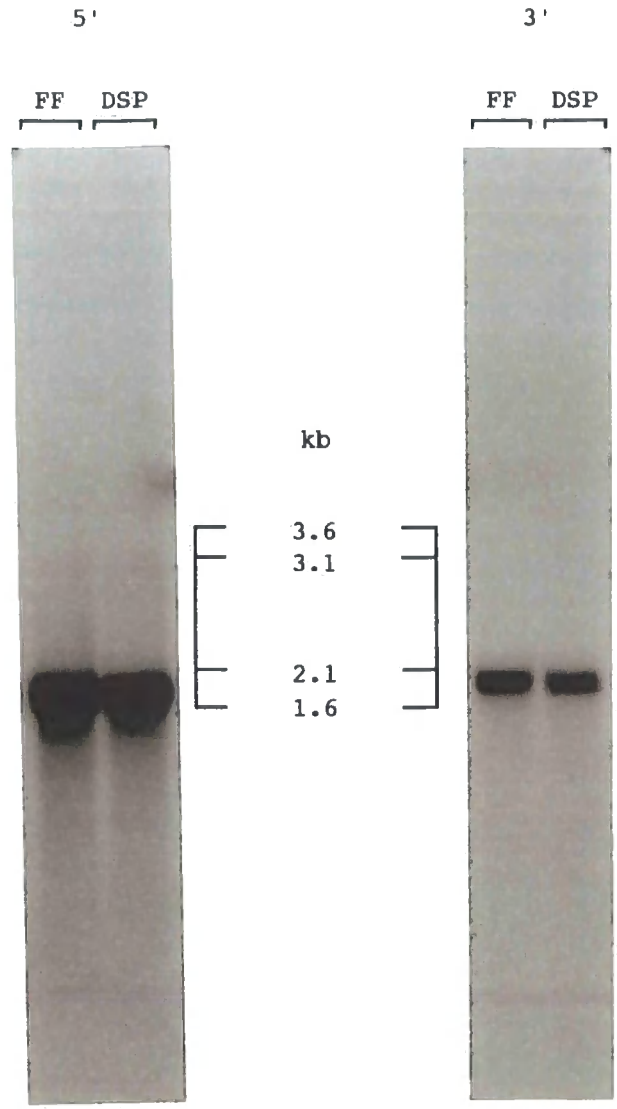


Fig. 17 Northern blot of RNA from Feltham First (FF) and Dark Skinned Perfection (DSP) cotyledons probed with fragments 5' and 3' of the *Bst* EII site in pLG1.63, vicilin 47k encoding cDNA. The position standard RNA ran to on the original gel is marked

CHAPTER SEVEN: VICILIN GENE, *vicJ*, SUBFAMILY - DISCUSSION

The *vicJ* sequence

As previously mentioned in the results section and can clearly be seen from the *vicJ* and cDNA comparison (fig. 14), the sequence of this gene diverges from that of the cDNA at a point 1745bp from the apparent start codon. The sequence therefore will be discussed in three parts; the vicilin 47k coding sequence (*vicJ* - up to the point of divergence and pLG1.63), the 5' flanking sequence, and the diverged sequence.

1) The vicilin 47k coding sequence

The predicted aa sequence of *vicJ* matches very closely that of the full length cDNA, pLG1.63 encoding the 47k Mr vicilin, with only 3 base differences and 2aa substitutions (fig 14). Between pLG 1.63 and the previously reported cDNAs, pDUB7 and pDUB4 (Lycett *et al.*,1983a), there is more variation (11 bases and 9aa) demonstrating that these cDNAs derive from separate, closely related genes. The gene encoding pLG1.63 will be referred to as *vicK* and that encoding pDUB7 and pDUB4 *vicL*.

The N-terminus of the 33k Mr vicilin subunit (which is derived from the vicilin 47k precursor) has been determined to be R-S-D and it was previously thought that the 47k vicilin precursor had a 15aa leader sequence (Lycett *et al.*,1983a). From the *vicJ* sequence and that of pLG1.63, it now appears that translation commences a further 27bp 5' and the cDNA extends beyond the sequence previously suggested as a possible cap site. The sequence flanking the further 5' start codon matches exactly the preferred translation start consensus A/C NNATGG, whereas YNNATGY was found to be predominantly non-functional (Kozak,1981).

A second argument in favour of the extended leader sequence is that this would result in a distance from the cap site, determined by S1 mapping, of around 50b to the first ATG, which falls within the

consensus distance (11-68b) (Messing *et al.*,1983), although with such a variable range 77b to the second ATG would not be totally implausible. A leader sequence of 24 residues compares well with those of the other pea storage proteins, legumin 21aa (Lycett *et al.*,1984a) and convicilin 28aa. The shorter leader sequence was however, sufficient to direct the product of a reporter gene to the inside of the endoplasmic reticulum when transformed into tobacco (Pang *et al.*,1992).

Surprisingly the cleavage point which results in the observed N-terminus is in the least favourable position, C-terminal to one of the three serine residues in this region of the protein, as judged by consensus patterns (von Heijne 1983). Removal of the signal peptide would result in a 414aa polypeptide with a Mr of 47,163 (based on the pLG1.63 sequence), this is in close agreement with the observed Mr for this protein of 47k (Gatehouse *et al.*,1981).

Some members of the vicilin gene family have been shown to undergo post-translational proteolysis (Gatehouse *et al.*,1981, Chrispeels *et al.*,1982). Two major, potential, proteolytic cleavage points have been identified, between the α, β and γ subunits (Gatehouse *et al.*,1982a & 1983). At the α:β cleavage point both *vicJ* and pLG1.63 predict RSLK:DRRQ which is not cleaved but at the β:γ cleavage point, both sequences predict GKEN:DKEE which should be cleaved (Gatehouse *et al.*,1984). Cleavage at this second site of the pLG1.63 encoded polypeptide would result in subunits of 306 and 108 residues with predicted Mrs of 35,109 and 12,072, again in close agreement with vicilin subunit observed Mrs (Gatehouse *et al.*,1981).

When the hydrophilicity profile for pLG1.63 encoded protein (fig. 18) is examined, it is obvious that as expected, the cleaved β:γ processing site (between residues 306 and 307) lies on a markedly hydrophilic region. This would predict that this region is located on

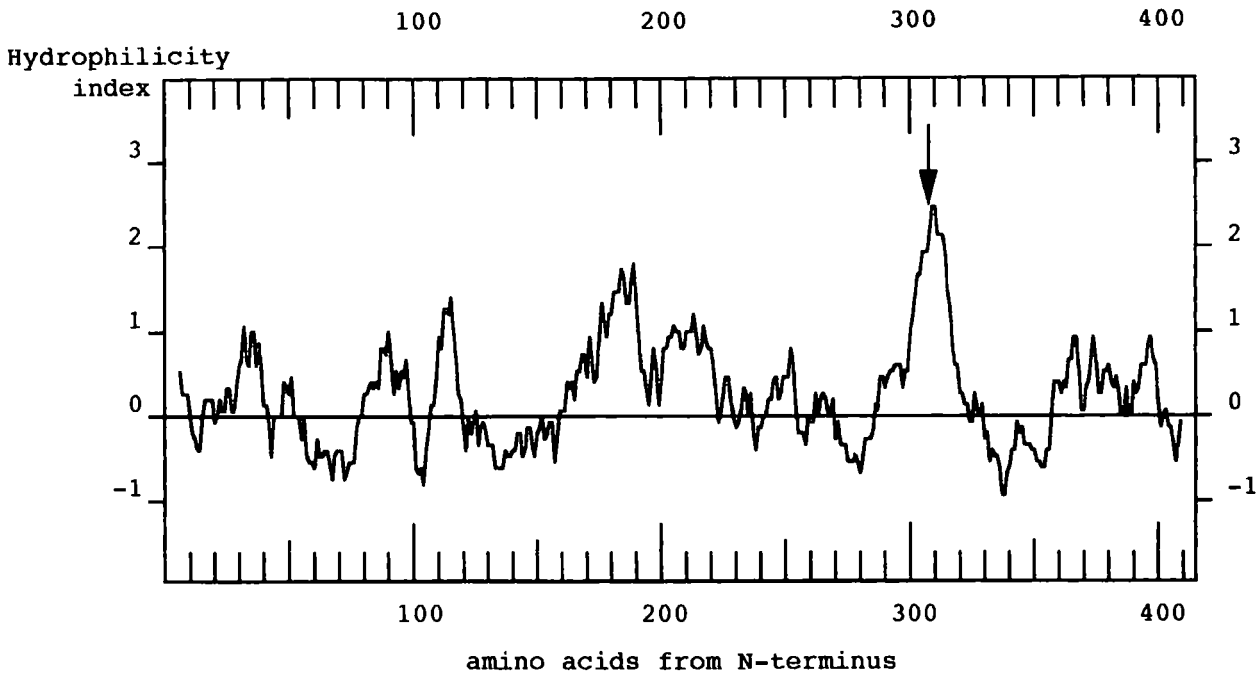


Fig. 18 Hopp Woods hydrophilicity plot for pLG1.63 encoded protein sequence (without signal peptide). The arrow shows the position of the β:γ subunit processing site between residues 306 and 307.

the outside of the vicilin molecule and is therefore susceptible to proteolysis. The sequence predicted by pLG1.63 retains the N-glycosylation site noted in the previous cDNA sequence (Lycett *et al.*,1983a) and suggests that this cDNA encodes the 16k rather than 12.5K vicilin γ subunit (Gatehouse *et al.*,1982a).

The position of the stop codon in pLG1.63 results in a shorter polypeptide (by 17aa) than that predicted by the previously sequenced 47k vicilin encoding cDNA (Lycett *et al.*,1983a). This predicted C-terminus for pLG1.63 is 5 residues upstream from the observed C-terminus of the 16k vicilin subunit (glycosylated 12.5K). These differences in the cDNAs cast doubts as to whether the variation in position between observed and predicted C-termini is due to C-terminal processing and suggests that this is due to variation in gene coding sequence length.

2) The 5' flanking sequence of *vicJ*

The sequence of 502bp 5' from the predicted start of translation in *vicJ* has been determined. As conserved sequences within the vicilin gene family and their roles as possible regulatory sequences within this region have been discussed with respect to convicilin (qv.), these will only be mentioned briefly here. Features will be discussed in succession working in a 5' direction from the start codon, they are noted on fig 14.

The S1 mapping experiment with *vicJ* gave a positive result and predicts a transcription start point within a region CATCATCT containing two consensus sequences CATC found at transcription start points in plant genes (Joshi 1987). A TATA box occurs at around 32bp 5' from the predicted transcription start region of *vicJ* and this is preceeded by a CCAAAT sequence, again of a conventional nature. The "vicilin box" in *vicJ* is also a good fit to the consensus (Gatehouse *et al.*,1986).

Although *vicJ* does not contain the two G+C rich regions further upstream from the "vicilin box" which are conserved between phaseolin, conglycinin and convicilin, a similar region, CCACCACCC, is found 39bp 5' from the "vicilin box" and this is conserved between *vicJ* and pea vicilin genes *vicB* (Boulter *et al.*,1990) and *Vc-4* (Higgins *et al.*,1988) and *Vicia faba* vicilin gene *Vfvic1* (Weschke *et al.*,1987). *VicJ* appears then to contain within its 5' flanking sequence those features so far associated with a functional legume seed storage protein gene (Gatehouse *et al.*,1986).

3) The diverged sequence of *vicJ*

Although the 5' sequence and coding sequence up to the point of divergence appear functional, can the same be said of the region at the 3' end? The sequence beyond the point of divergence continues to code

for a further 14 residues before a stop codon is reached. The residues encoded are not atypical when compared to the rest of the sequence and, although the nucleotide sequence is rather A+T rich (71% A+T), similar regions exist within the vicilin encoding sequence (at the 5' ends of exons 3 and 5, for example).

Beyond the stop codon the nucleotide sequence continues with the sort of composition which would not be remarkable at the 3' end of an mRNA and beginning at 125,150,185 and 210bp beyond the stop codon are sequences which conform to the plant consensus polyadenylation signal G/A $ATAA_{(1-3)}$ (Messing *et al.*,1983). Although there is no sequence homology between the 3' non-translated sequences of *vicJ* and pLG1.63, there appears to be no reason, so far as is known, why this should prevent transcription and expression of *vicJ*.

To determine the nature of the rearrangement at the 3' end of *vicJ*, the genomic clone was probed with the 3' end of pLG1.63. No homologous sequence could be detected, indicating that either a deletion of this sequence has occurred, or else an insertion of more than 5.6kbp (the distance from the point of divergence to the 3' end of the genomic clone). Large insertions have been shown to have occurred in plants, such as those of 17 and 30kbp due to the insertion of transposable elements in *Antirrhinum* and maize (Bonas *et al.*,1984, Doring *et al.*,1984). Due to the number of genes estimated to be present in the vicilin 47k gene subfamily, it is impossible to tell from fragments hybridised to by the 3' end of the cDNA probe, on the genomic blot (fig 15b), whether there is a fragment present in the genome corresponding to that missing from λJC3.

It is possible that the diverged sequence forms an extra intron which includes the nucleotide before divergence and so starts GT (see fig. 14) but this seems unlikely. Firstly, no other expressed genes of

the vicilin family so far sequenced have deviated from the "normal" intron position and number. Secondly, if this were an intron, just as if it were an inserted sequence, it must be greater than 5.6kbp in length, an exceptional size for an intron (Nasra & Deacon, 1982). Finally, the nucleotide composition, although A+T rich (68%), is not typical of that of a plant intron, especially the G+A rich region 300-500bp from the point of divergence.

As has been stated above, large insertions are known to have occurred as the result of transposition and it is possible that this is the case in *vicJ*. Insertions presumed to have resulted from the action of transposable elements have been found in or near pea genes (Shirsat, 1988, Bhattacharyya *et al.*,1990). Without the original 3' end of this gene (if present), it is not possible to look for the repeats around the insertion sites which are symptomatic of transposable elements (Freeling 1984).

When the diverged sequence on *vicJ* was used to probe genomic DNA (fig. 15c), it hybridised to multiple bands at a high level of intensity indicating that it's sequence is repeated many times in the pea genome. Such repetitive sequences had previously been noted on two *Eco* RI fragments from this genomic clone, including the one containing *vicJ*, and it was estimated that the levels of this repeated sequence within the pea genome were similar to that of ribosomal genes - around 0.15% (Ellis *et al.*,1986) (equivalent to 7,500 copies of a 1kbp fragment in the pea genome). When λJC3 was probed with the diverged sequence from *vicJ*, only the fragments containing *vicJ* were hybridised to, demonstrating that the other repeat sequence present is different to that in *vicJ*. The terminal regions of transposable elements are found repeated in the genome but usually only up to 50 copies are found (Nevers *et al.*,1986).

The vicilin 47k gene subfamily

Probes from the 47k gene subfamily have been shown not to cross-hybridise with those from other types of vicilin genes at the stringency used to wash the genomic blots (Ellis *et al.*,1986). Genomic blots probed with the vicilin encoding region of *vicJ* and pLG1.63 (fig 15) demonstrate the presence of several members within this subfamily and agree with the previous estimate of five to seven members (Domoney & Casey, 1985). Bands equivalent to those predicted by λJC3 are present in both DSP and FF lines, therefore; a) the divergence in *vicJ* is not due to rearrangement during cloning, and b) *vicJ* is present in the FF genome in addition to the definitely functional genes *vicK* and *vicL* encoding the cDNAs from this line.

Expression of *vicJ*

In the S1 mapping experiment, the fragment from *vicJ* used is identical to pLG1.63 as far as the cDNA extends. Therefore the positive result is not surprising and cannot be taken as indicating that *vicJ* itself is expressed but presumably shows the transcription start point of the gene encoding pLG1.63 *(vicK)*. The less intense, smaller bands seen on the S1 mapping gel may be due to protection by the mRNA from related genes. These may have sequence mismatches which would result in the nuclease nicking the DNA strand between the cap site and the labelled end, so producing shorter labelled fragments.

As stated above there appears no reason why the *vicJ* gene should not be transcribed in a truncated form, translated and processed. However, no truncated mRNA could be detected with the cDNA pLG1.63 in cotyledon RNA in which the "normal" mRNA was being expressed (fig 17). It has generally been found that mRNA does not accumulate from genes which contain deleterious mutations (Vodkin *et al.*,1983, Voelker *et*

al.,1990, Thompson *et al.*,1991), although the mechanism which prevents this is unknown. A transposable element induced mutation of the maize sucrose synthase gene however, results in a shortenened mRNA being transcribed and a mutant protein expressed (Federoff *et al.*,1983).

Some features of the mutation of the starch branching enzyme in wrinkled seeded peas (Bhattacharyya *et al.*,1990) parallel that of *vicJ*. In wrinkled seeded varieties the gene encoding the starch branching enzyme carries an insertion (with respect to the normal gene in round seeded lines) in the 3' end of the coding sequence. Like the 3' end of *vicJ*, the inserted sequence in this gene is repeated many times within the pea genome. However, unlike *vicJ*, mutant mRNA is transcribed, accumulating at lower levels than normal length mRNA. The termini of this inserted sequence are homologous to those of *Ac*-like transposons but no homology could be found between the beginning of the *vicJ* inserted sequence and the inverted repeat sequences of *Ac* or any other plant transposons (Vodkin, 1989).

Restriction mapping and sequencing of *legK*

A partial sequence for *legK* had previously been determined (Gatehouse *et al.*,1988) from a genomic clone, λJC5, isolated from DSP pea plants (Domoney *et al.*,1986a). A genomic clone containing a complete *legK* gene had also been isolated from FF pea plants and a *Sal* I subclone, pNY2, containing this gene produced (Yaish, 1990).

Restrictions were performed on this genomic subclone and from these a map of the sites was deduced (fig. 20). This *Sau*3A - *Sal* I subclone is 12.3 kbp in length and from the previously published restriction map of λJC5 (Gatehouse *et al.*,1988) the position of the *legK* coding sequence was established. Overlapping *Eco* RV and *Sph* I fragments were isolated, subcloned into pUC18 and 19 respectively and restriction mapped (fig. 20).

Fragments from these two clones were subcloned into M13 and automated DNA sequencing performed on them. The 3.4 kbp *Sph* I to *Sph* I region was sequenced in both directions with all restriction sites sequenced through. The sequence (fig. 21) is 3392bp in length, 1012bp lie 5' of the position of the start codon in *legJ* and 700bp were sequenced 3' of the stop codon. Over the 1533bp region of overlap between the previously determined *legK* sequence from DSP and this new sequence from FF there were no differences.

Comparison of the *legK* sequence with that of *legJ*

legK contains a 1500bp (500aa) ORF homologous to that of *legJ* (Gatehouse *et al.*,1988), however, at the expected N-terminus, *legK* lacks a methionine residue. This is due to an A to G substitution resulting in GTG, coding for valine, rather than ATG. When the *legK* sequence 3' from the TATA box is inspected, the first ATG codon is at base 117 on fig. 21

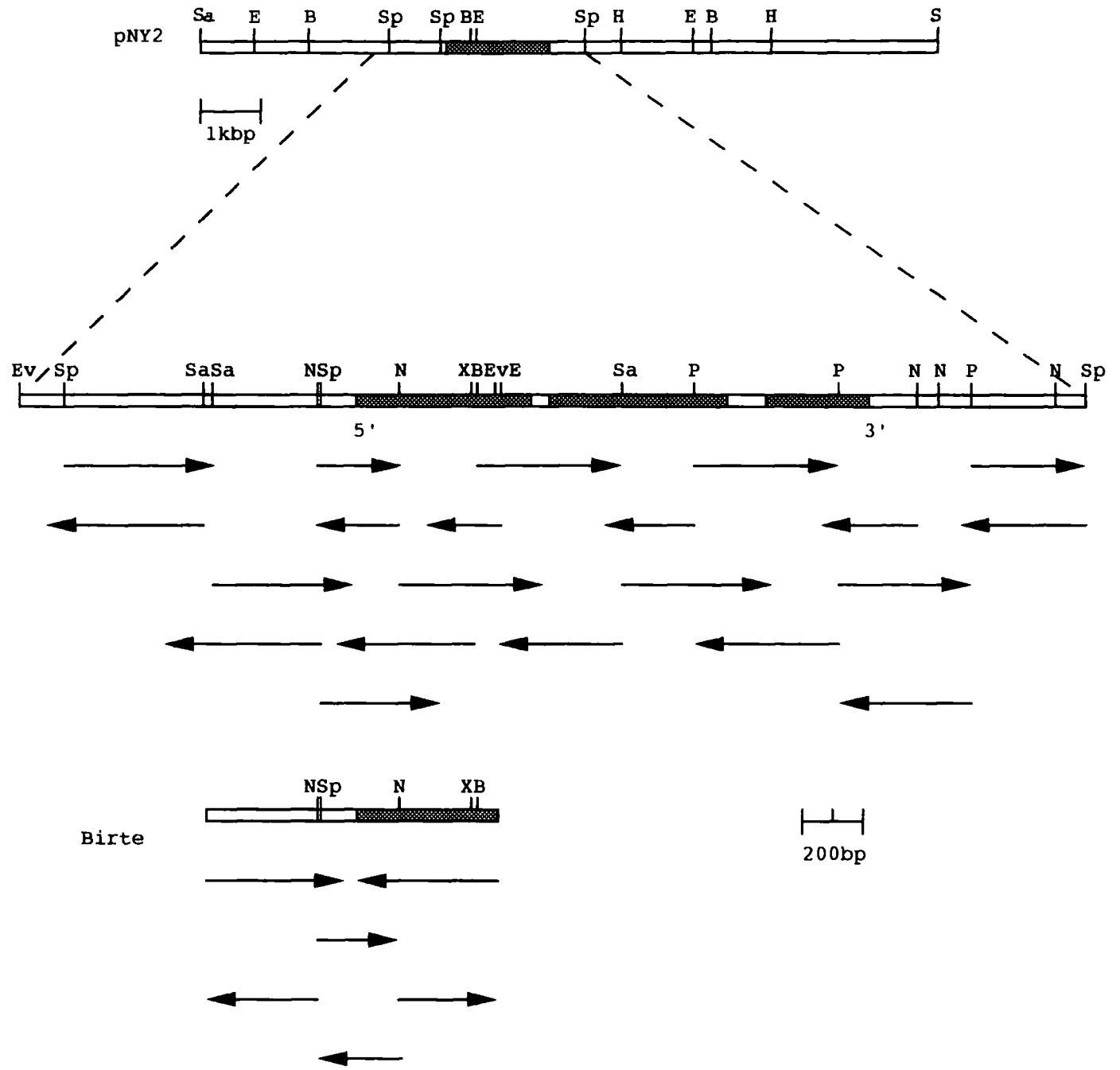
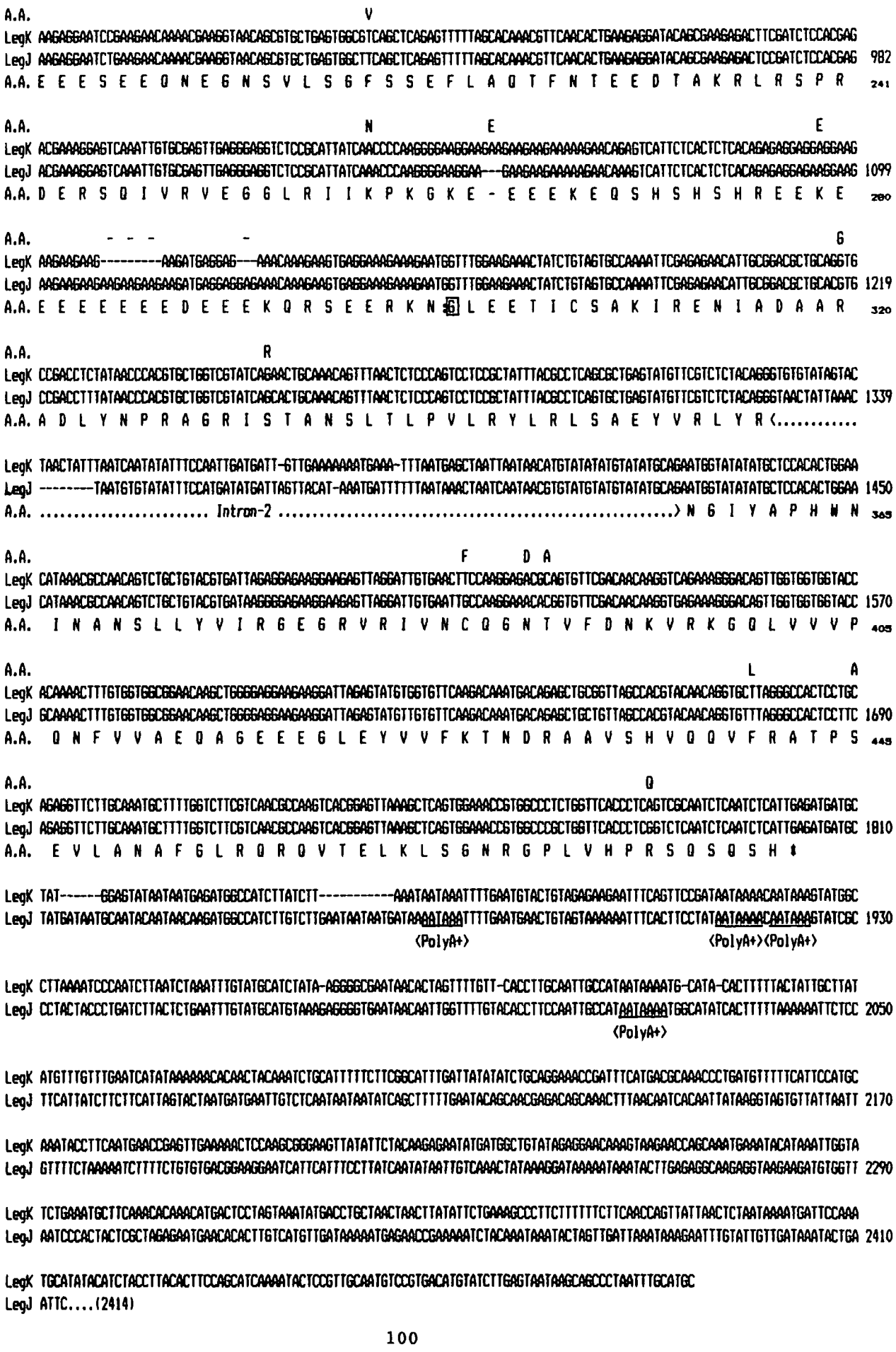


Fig. 20 Restriction map of the genomic subclone pNY2 from FF (above), the region sequenced from it (centre) and the PCR fragment from Birte (below) amplified using primers derived from the sequence. Restriction enzymes are abbreviated as: B, *Bam* HI; E, *Eco* RI; Ev, *Eco* RV; H, *Hind* III; N, *Nsi* I; P, *Pst* I; S, *Sal* I; Sa, *Sau* 3A; Sp, *Sph* I; X, *Xba* I. The coding sequence is highlighted and arrows represent individual sequencing runs.

Fig. 21 (Overleaf) Nucleotide sequence of *legK* from line Feltham First and *legJ* (from line DSP), the derived amino acid sequence of *legJ* is presented below the nucleotide sequences and residues differing from this in the *legK* encoded sequence are shown above. The subunit N-terminal residues are boxed and the cleavage sites marked (:). Consensus sequences in the 5' flanking region and polyadenylation signals are underlined The sequences from which the oligonucleotides were synthesised for amplification of *legK* in Birte are boxed. The transcription start site in legJ is arrowed.


A.A. V
LegK AAGAGGAATCCGAAGAACAAAACGAAGGTAACAGCGTGCTGAGTGGCGTCAGCTCAGAGTTTTTAGCACAAACGTTCAACACTGAAGAGGATACAGCGAAGAGACTTCGATCTCCACGAG
LegJ AAGAGGAATCTGAAGAACAAAACGAAGGTAACAGCGTGCTGAGTGGCTTCAGCTCAGAGTTTTTAGCACAAACGTTCAACACTGAAGAGGATACAGCGAAGAGACTCCGATCTCCACGAG 982
A.A. E E E S E E Q N E G N S V L S G F S S E F L A Q T F N T E E D T A K R L R S P R 241
A.A. N E E
LegK ACGAAAGGAGTCAAATTGTGCGAGTTGAGGGAGGTCTCCGCATTATCAACCCCAAGGGGAAGGAAGAAGAAGAAAAAGAACAGAGTCATTCTCACTCTCACAGAGAGGAGGAGGAAG
LegJ ACGAAAGGAGTCAAATTGTGCGAGTTGAGGGAGGTCTCCGCATTATCAAACCCAAGGGGAAGGAA---GAAGAAGAAAAAGAACAAAGTCATTCTCACTCTCACAGAGAGGAGAAGGAAG 1099
A.A. D E R S Q I V R V E G G L R I I K P K G K E - E E E K E Q S H S H S H R E E K E 280
A.A. G
LegK AAGAAGAAG---------AAGATGAGGAG---AAACAAAGAAGTGAGGAAAGAAAGAATGGTTTGGAAGAAACTATCTGTAGTGCCAAAATTCGAGAGAACATTGCGGACGCTGCAGGTG
LegJ AAGAAGAAGAAGAAGAAGAAGATGAGGAGGAGAAACAAAGAAGTGAGGAAAGAAAGAATGGTTTGGAAGAAACTATCTGTAGTGCCAAAATTCGAGAGAACATTGCGGACGCTGCACGTG 1219
A.A. E E E E E E E D E E E K Q R S E E R K N S L E E T I C S A K I R E N I A D A A R 320
A.A. R
LegK CCGACCTCTATAACCCACGTGCTGGTCGTATCAGAACTGCAAACAGTTTAACTCTCCCAGTCCTCCGCTATTTACGCCTCAGCGCTGAGTATGTTCGTCTCTACAGGGTGTGTATAGTAC
LegJ CCGACCTTTATAACCCACGTGCTGGTCGTATCAGCACTGCAAACAGTTTAACTCTCCCAGTCCTCCGCTATTTACGCCTCAGTGCTGAGTATGTTCGTCTCTACAGGGTAACTATTAAAC 1339
A.A. A D L Y N P R A G R I S T A N S L T L P V L R Y L R L S A E Y V R L Y R (.....
LegK TAACTATTTAATCAATATATTTCCAATTGATGATT-GTTGAAAAAAATGAAA-TTTAATGAGCTAATTAATAACATGTATATATGTATATGCAGAATGGTATATATGCTCCACACTGGAA
LegJ -------TAATGTGTATATTTCCATGATATGATTAGTTACAT-AAATGATTTTTTAATAAACTAATCAATAACGTGTATGTATGTATATGCAGAATGGTATATATGCTCCACACTGGAA 1450
A.A. ..... Intron-2 .....) N G I Y A P H W N 365
A.A. F D A
LegK CATAAACGCCAACAGTCTGCTGTACGTGATTAGAGGAGAAGGAAGAGTTAGGATTGTGAACTTCCAAGGAGACGCAGTGTTCGACAACAAGGTCAGAAAGGGACAGTTGGTGGTGGTACC
LegJ CATAAACGCCAACAGTCTGCTGTACGTGATAAGGGGAGAAGGAAGAGTTAGGATTGTGAATTGCCAAGGAAACACGGTGTTCGACAACAAGGTGAGAAAGGGACAGTTGGTGGTGGTACC 1570
A.A. I N A N S L L Y V I R G E G R V R I V N C Q G N T V F D N K V R K G Q L V V V P 405
A.A. L A
LegK ACAAAACTTTGTGGTGGCGGAACAAGCTGGGGAGGAAGAAGGATTAGAGTATGTGGTGTTCAAGACAAATGACAGAGCTGCGGTTAGCCACGTACAACAGGTGCTTAGGGCCACTCCTGC
LegJ GCAAAACTTTGTGGTGGCGGAACAAGCTGGGGAGGAAGAAGGATTAGAGTATGTTGTGTTCAAGACAAATGACAGAGCTGCTGTTAGCCACGTACAACAGGTGTTTAGGGCCACTCCTTC 1690
A.A. Q N F V V A E Q A G E E E G L E Y V V F K T N D R A A V S H V Q Q V F R A T P S 445
A.A. Q
LegK AGAGGTTCTTGCAAATGCTTTTGGTCTTCGTCAACGCCAAGTCACGGAGTTAAAGCTCAGTGGAAACCGTGGCCCTCTGGTTCACCCTCAGTCGCAATCTCAATCTCATTGAGATGATGC
LegJ AGAGGTTCTTGCAAATGCTTTTGGTCTTCGTCAACGCCAAGTCACGGAGTTAAAGCTCAGTGGAAACCGTGGCCCGCTGGTTCACCCTCGGTCTCAATCTCAATCTCATTGAGATGATGC 1810
A.A. E V L A N A F G L R Q R Q V T E L K L S G N R G P L V H P R S Q S Q S H *
LegK TAT------GGAGTATAATAATGAGATGGCCATCTTATCTT----------AAATAATAAATTTTGAATGTACTGTAGAGAAGAATTTCAGTTCCGATAATAAAACAATAAAGTATGGC
LegJ TATGATAATGCAATACAATAACAAGATGGCCATCTTGTCTTGAATAATAATGATAAAATAAATTTTGAATGAACTGTAGTAAAAAATTTCACTTCCTATAATAAAACAATAAAGTATCGC 1930
(PolyA+) (PolyA+)(PolyA+)
LegK CTTAAAATCCCAATCTTAATCTAAATTTGTATGCATCTATA-AGGGGCGAATAACACTAGTTTTGTT-CACCTTGCAATTGCCATAATAAAATG-CATA-CACTTTTTACTATTGCTTAT
LegJ CCTACTACCCTGATCTTACTCTGAATTTGTATGCATGTAAAGAGGGGTGAATAACAATTGGTTTTGTACACCTTCCAATTGCCATAATAAAATGGCATATCACTTTTTAAAAAATTCTCC 2050
(PolyA+)
LegK ATGTTTGTTTGAATCATATAAAAAACACAACTACAAATCTGCATTTTTCTTCGGCATTTGATTATATATCTGCAGGAAACCGATTTCATGACGCAAACCCTGATGTTTTTCATTCCATGC
LegJ TTCATTATCTTCTTCATTAGTACTAATGATGAATTGTCTCAATAATAATATCAGCTTTTTGAATACAGCAACGAGACAGCAAACTTTAACAATCACAATTATAAGGTAGTGTTATTAATT 2170
LegK AAATACCTTCAATGAACCGAGTTGAAAAACTCCAAGCGGGAAGTTATATTCTACAAGAGAATATGATGGCTGTATAGAGGAACAAAGTAAGAACCAGCAAATGAAATACATAAATTGGTA
LegJ GTTTTCTAAAAATCTTTTCTGTGTGACGGAAGGAATCATTCATTTCCTTATCAATATAATTGTCAAACTATAAAGGATAAAAATAAATACTTGAGAGGCAAGAGGTAAGAAGATGTGGTT 2290
LegK TCTGAAATGCTTCAAACACAAACATGACTCCTAGTAAATATGACCTGCTAACTAACTTATATTCTGAAAGCCCTTCTTTTTTCTTCAACCAGTTATTAACTCTAATAAAATGATTCCAAA
LegJ AATCCCACTACTCGCTAGAGAATGAACACACTTGTCATGTTGATAAAAATGAGAACCGAAAAATCTACAAATAAATACTAGTTGATTAAATAAAGAATTTGTATTGTTGATAAATACTGA 2410
LegK TGCATATACATCTACCTTACACTTCCAGCATCAAAATACTCCGTTGCAATGTCCGTGACATGTATCTTGAGTAATAAGCAGCCCTAATTTGCATGC
LegJ ATTC....(2414)

56bp downstream from the start of *legJ*. This ATG codon is out of frame with the *legJ* sequence and is followed after five codons by a TGA stop codon.

The remaining *legK* coding sequence is highly homologous to *legJ* with only 45 more nucleotide changes (in the 1500bp of *legK*) resulting in 20 further aa substitutions. There are also four codon deletions and one addition to *legK* with respect to *legJ*. The two intron positions are conserved between the two genes, although intron 1 is 56bp shorter in *legK*, as a result of two deletions and intron 2 carries an 8bp addition with respect to *legJ*.

As can be seen from fig. 21, there is considerable homology between the 5' flanking regions of the two genes to the full extent of the sequence, although both contain considerable deletions or insertions with respect to each other. At the 3' end of the genes homology extends for about 250bp from the stop codons. After a point 16bp 3' of the fourth polyadenylation signal, the sequences diverge completely.

Isolation and sequencing of the 5' region of *legK* from Birte

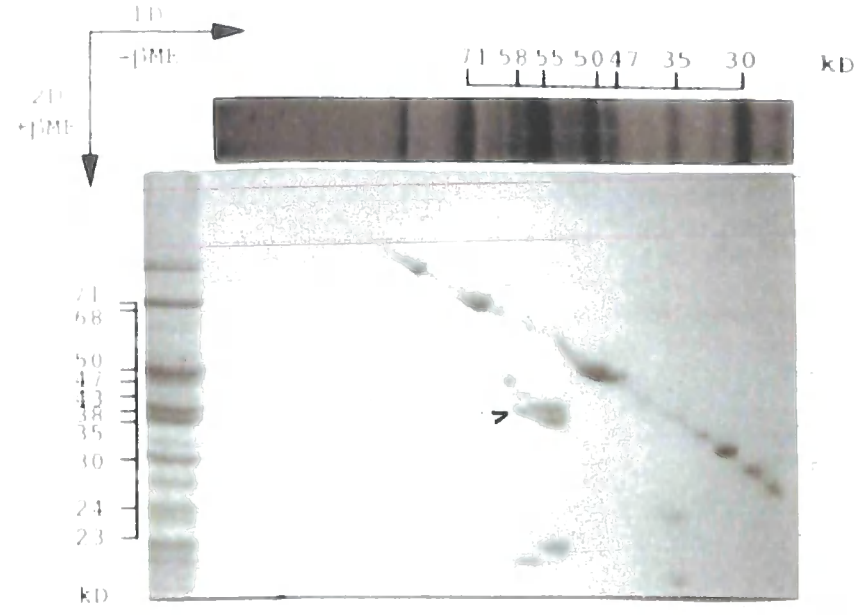
A cDNA had been isolated from Birte pea cotyledons which was identical to *legK* except for 2bp in the 910bp cDNA (Domoney *et al.*,1986b, Gatehouse *et al.*,1988). It was assumed then that the cv Birte contains a funtional *legK* gene. To test this hypothesis, the 5' region of *legK* was amplified from Birte DNA using the polymerase chain reaction (PCR). Oligonucleotide primers were synthesised using the determined *legK* sequence, their positions are shown on fig. 21. The 5' primer was synthesised using sequence specific to *legK* (not found in *legJ*) around 530bp from the start codon position. The 3' primer was made complimentary to the *legK* sequence around 400bp 3' from the same point, it included in its centre a single base substitution with respect to

legJ.

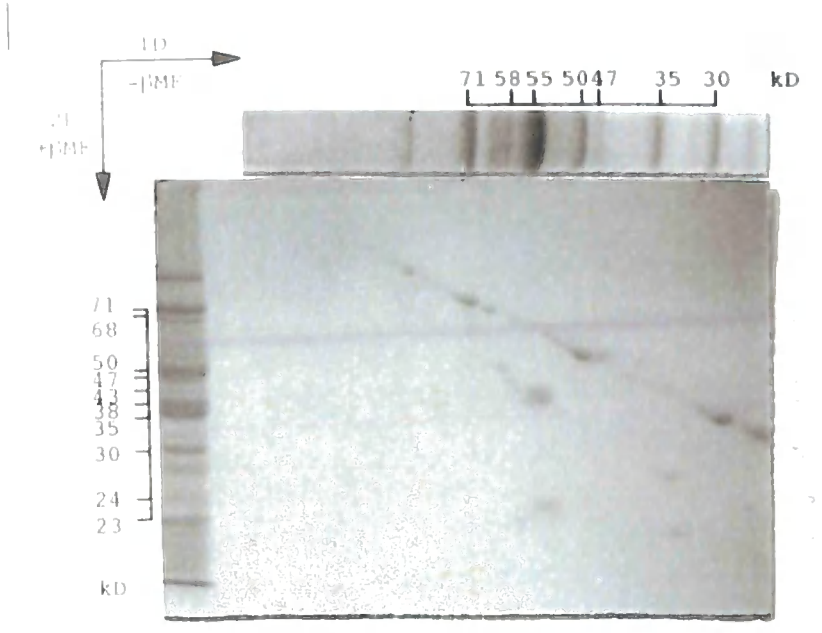
Birte genomic DNA was amplified using these primers, the product electrophoresed through agarose gel, and the DNA isolated by electroelution. The ends of the amplified product were "polished" using T4 polymerase and cloned into pUC18 vector. Automated sequencing was performed directly on this plasmid, in both orientations, and subsequently on subclones produced from this. The insert was sequenced fully in both directions and all restriction sites were sequenced through (fig. 20). The sequence extends through both primers and is identical to that of the FF *legK* sequence except for the codon at residue 1 of the *legJ* sequence which in this case is an ATG start codon.

Identification of protein subunits corresponding to *legK*

To try and identify the legumin species encoded by *legK*, which should be present in an extract from FF seeds but not from those of Birte, two dimensional polyacrylamide gels were run. Extracts of mature seeds were electrophoresed in the first dimension under non-reducing conditions and then in the second dimension under reducing conditions. Fig. 22 shows the resultant gels, it can be seen that the FF extract lacks a subunit, which is present in the Birte extract, of higher non-reduced Mr than the major legumin α-subunits. The β-subunits are not so clearly resolved but the subunit of equal mobility in the first dimension to that of the α-subunit present only in Birte, is much reduced in intensity in FF compared to Birte.



Birte



Feltham First

Fig. 22 Two dimensional polyacrylamide gels of seed extracts from pea lines Birte and Feltham First (FF). Gels were run in the first dimension (1D) under non-reducing conditions and in the second dimension (2D) under reducing conditions. The Mr of polypeptides in the seed extracts run under both conditions is indicated. The arrow marks the legumin α-subunit present in Birte but not FF.

The functional *legK* gene in Birte

From the results of the PCR experiment on DNA from pea line Birte, it can be seen that this line contains a functional *legK* gene. This conclusion is supported by the presence of a cDNA identical to *legK* (except for 2bp in 907)(Domoney *et al.*,1986b Gatehouse *et al.*,1988) and an extra legumin protein subunit pair present in Birte compared to FF (see below). The start codon is in the same position as that of the other gene sequenced from this subfamily, *legJ* (fig. 21). Although this sequenced fragment from Birte was amplified by PCR and errors might be expected in the faithfullness of copying by *Taq* polymerase (Karlovsky, 1990), none were obvious, as the sequence was identical to that of *legK* from FF which had been subcloned by conventional means. It seems highly unlikely that the only error in 945bp should be the return of the functional start codon!

Comparison with *legJ*

The coding sequence of *legK* is very similar to that of *legJ* which has previously been shown to encode a minor legumin subunit pair (Gatehouse *et al.*,1988). There is close homology at the nucleotide level (97%) which is mirrorred at the aa level by only 20 aa substitutions, with 3 deletions and one addition in *legK* compared to *legJ*. This homology accounts for the detection of *legJ* mRNAs by *legK* coding sequence probe (Croy *et al.*,1988 Thompson *et al.*,1991).

The non-coding sequence of the two genes is also well conserved, the introns are 96% and 74% homologous if deletions are ignored and suggest a relatively recent date for the divergence of the two genes. The 3' flanking sequence is well conserved for around 250bp and includes the four polyadenylation signals found in *legJ*. This would suggest that

legK derived mRNAs terminate at differing points as is the case with *legJ*, *legL* (Thompson *et al.*,1991) and *legS*$_{\lambda}$ (pCD32) (Domoney *et al.*,1986b, Gatehouse,JA., unpublished results).

The 5' flanking sequences show homology over the entire region sequenced (600bp from *legJ*), despite the considerable deletions in both with respect to the other. Both genes contain the conserved "legumin box" although there is one base mismatch in the less well conserved 3' half (Gatehouse *et al.*,1986). The "legumin box" includes the RY repeats found in many legume seed-protein genes (Dickinson *et al.*,1988), in *legJ* and *legK* they take the usual form, CATGCATG, compared with CATGCAAG found in this region of the *legA* subfamily (Lycett *et al.*,1985). The *legJ* and *legK* sequences lack the "core enhancer" sequence (CCGCCACC) found in the 5' regions of *legA* type legumin genes and which forms part of the "vicilin box" of legume vicilin genes (Gatehouse *et al.*,1986). What bearing these last two differences have on the differential expression of these two legumin subfamilies is not clear at present.

The mutation of *legK* in Feltham First

In the FF pea line, *legK* carries a mutated start codon (ATG to GTG). This is the only sequence difference found in the *legK* genes in the lines FF, DSP and Birte in the region where they overlap each other, except for the 2bp differences between the Birte cDNA and *legK* from DSP and FF. This confirms the very recent nature of the *legK* mutation which must have occurred after the separation of FF and Birte parent lines. Although *legK* is now a pseudogene in line FF, it has not accumulated further mutations, in contrast to another legumin pseudogene, *legD*, found in the *legA* gene subfamily. *LegD* has in-frame stop codons but also exhibits many other differences when compared to functional genes from this subfamily (Bown *et al.*,1985). No mRNA of a size corresponding to

that predicted from *legD* could be detected when other members of its subfamily were being expressed. The mutation in *legK* resembles that found in a null allele of a soybean legumin gene, glycinin Gy_4, where a similar mutation (ATG to ATA) of the start codon occurs (Scallon *et al.*,1987). This mutation in Gy_4 is accompanied by only two other nucleotide changes in 3.5kbp sequenced. In Gy_4, as in *legK*, there is another potential start codon (56bp) further 3' to the mutated one, but this is out of frame and encodes a reading frame of only 6 residues.

Hybridisation of a *legK* specific oligonucleotide to mRNA from FF failed to detect any transcripts from this gene *in vivo* (Thompson *et al.*,1991). This work demonstrates that translationally ineffective mRNAs fail to accumulate, presumably due to their failure to remain bound to polysomes. Similar results were obtained with the glycinin mutant gene (see above) where very low levels of mRNA could be detected, but not attached to polysomes. A mutation of a *Phaseolus* lectin gene, causing a single base deletion, results in in-frame stop codons. mRNA levels from this are also drastically reduced, although the gene's promoter region was shown to be functional when transformed with a reporter gene into tobacco (Voelker *et al.*,1990).

There is a 6bp repeat in the 5' flanking region of *legK*, which is not present in *legJ*, this lies 20bp from the start codon position. Before this region of *legK* had been sequenced from Birte it was suggested that this repeat might affect mRNA stability, as it was located in the area where ribosomal binding would be expected (Thompson *et al.*,1991). However, now that this repeat has been found in the functional *legK* gene from Birte, it is shown that this cannot be the case. This also rules out any connection between occurrence of this repeat and the mutation of the start codon.

The *legK* phenotype

When protein from FF and Birte seeds was analysed by two dimensional gel electrophoresis it could be seen that Birte contained an extra minor legumin subunit pair. Although these gels were run using total extract the subunit pattern was essentially the same as that obtained using purified legumin (Matta *et al.*,1981) except that the α-subunits (labelled *legJ* and *legX* on fig. 23) appear to have been resolved on these gels (fig. 22). The subunit pair L2 (Matta *et al.*,1981) identified with *legJ* (Gatehouse *et al.*,1988) was present in both extracts but the subunits present in Birte and not in FF cannot be those tentatively ascribed to *legK* by Croy *et al.* (1988). The figure (fig. 23) these authors used shows purified legumin from FF seeds and so does not contain the *legK* subunit pair. Possibly the subunit pair ascribed to *legK* could be the product of another member of this subfamily (see below). The presence and expression of a third member of this subfamily, *legL*, can be deduced by the number of bands hybridised to on genomic blot by *legJ* subfamily specific probe (Domoney & Casey, 1985, Levasseur, 1988) and by FF mRNAs hybridised to by *legK* coding sequence probes, but unaccounted for by *legJ* specific probes (Thompson *et al.*,1991).

The α-subunit of *legK* appears on gel to be of considerably lower Mr than that of the subunit pair ascribed to *legJ* (fig. 23, approximately 37.5k and 43k, respectively, using the size estimates of Matta *et al.*,1981). The *legK* sequence predicts a subunit Mr of 33,889 compared to 34,485 for *legJ* α-subunit (Gatehouse *et al.*,1988), which does not account for the large size difference on gel. The reason for the size discrepancy is not clear. It has been shown that legumin is not glycosylated (Gatehouse *et al.*,1980), so differential carbohydrate attachment cannot be the cause. It may be that those subunits ascribed

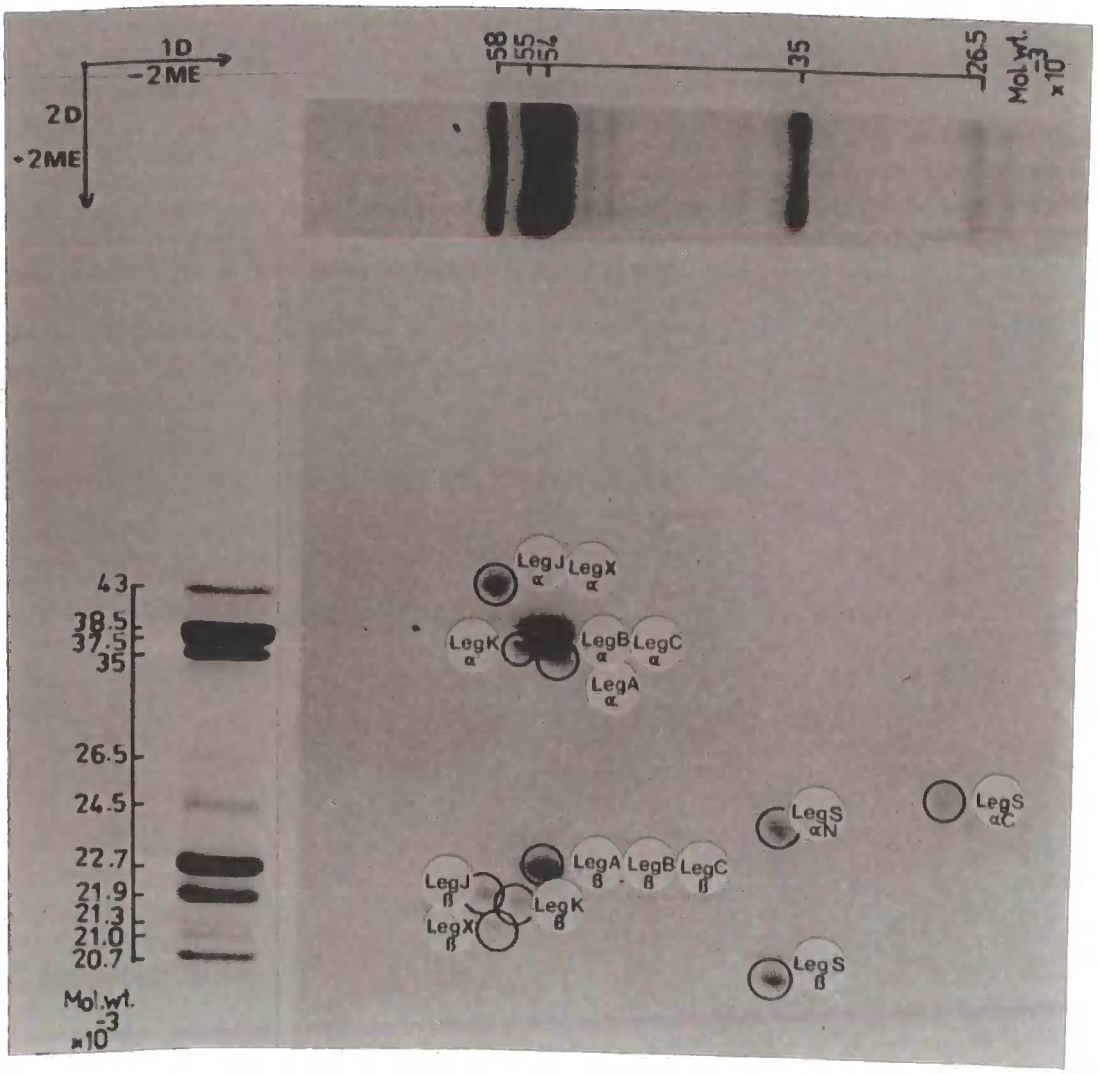


Fig. 23 Two dimensional polyacrylamide gel of purified legumin from Feltham First, run in the first dimen ion under non-reducing conditions and in the second, under reducing conditions. The Mr of legumin polypeptides run separately under both conditions is indicated. Taken from Croy *et al.* (1988).

to *legJ* are in fact the product of *legL*, whose homology to *legJ* would be high, accounting for the agreement of the peptide sequences from these subunits with that derived from *legJ*, but which could contain a duplication (or duplications) in the variable C-terminal region of the α-subunit, as is the case in the *legA* gene subfamily (Lycett *et al.*,1984b) and *legS* (Domoney *et al.*,1986b). This hypothesis is supported by two results. 1) a 2D isoelectric focusing gel of legumin α-subunits (Krishna *et al.*,1979) which shows *legS* subunits and the subunit previously ascribed to *legJ* (fig. 23) to have lower pIs than the main group of legumin subunits; the enlargement by duplication of the acidic region at the C-terminal of these subunits would account for the lower pI values. 2) mRNA encoding *legL* was shown to have a significantly higher Mr than that encoding *legJ* (Thompson *et al.*,1991). If the spot on fig. 23 labelled *legJ* is, however *legL*, then that labelled *legK* is presumably *legJ* and is of a slightly lower Mr than that shown to be *legK* in Birte. This is a reversal of the relative sizes predicted by the sequences, but the size differences are small, and may reflect the slightly changed amino acid composition of *legK* with respect to *legJ*.

The legumin gene family in Feltham First

This work and other recent studies have shed light on the nature of the legumin gene family and a final picture can be drawn, although further additions of distantly related genes cannot be ruled out. The variability between pea lines and varieties (Casey, 1979, Matta & Gatehouse, 1982) means that these specific comments can only be applied to the FF line, but the main framework of the family should be applicable to other lines. Fig. 24 shows the outline of the legumin family tree based on the homology between the N-terminal sequences of both subunits (all the available sequence from *legX*).

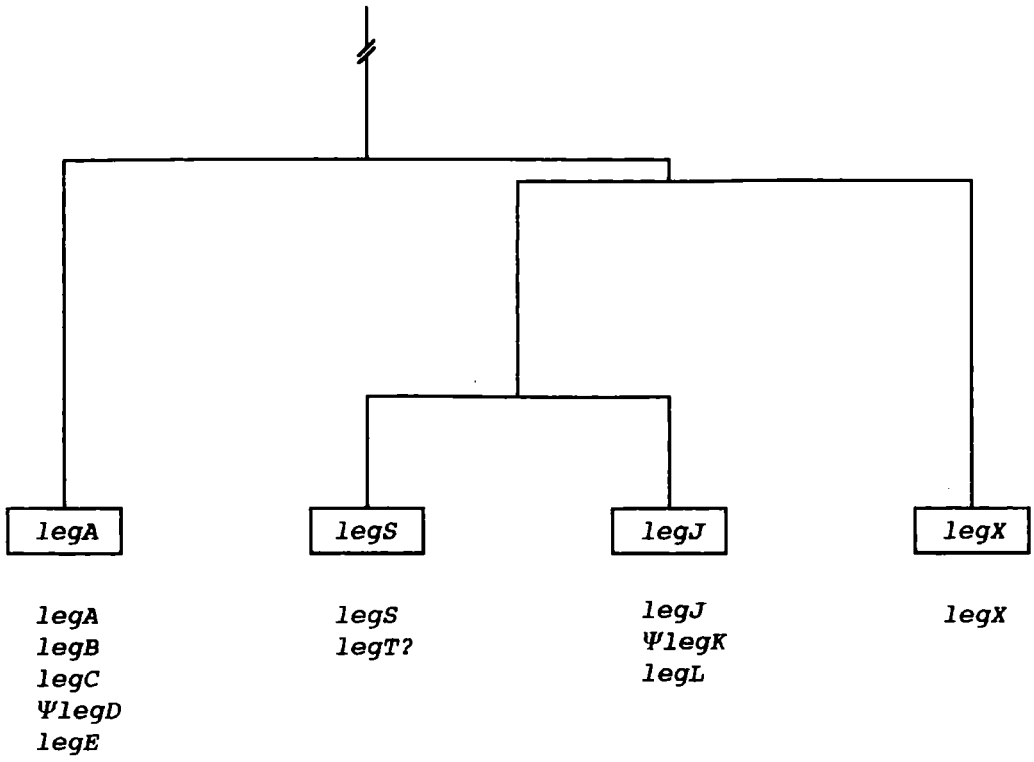


Fig. 24 Legumin gene family tree - diagram demonstrating the relationship between, and composition of, the legumin gene subfamilies in Feltham First. The distance between the branch points in the vertical axis is based on the mean identities between the N-terminal sequences of the subunits encoded by the subfamily members. Sequences used extend as far as the sequence determined on the *legX* encoded protein subunits (March *et al.*,1988). Position and spacing in the horizontal axis is arbitrary.

The *legA* subfamily consists of five members lying on four *Eco* RI fragments (Croy *et al.*,1982, Domoney & Casey, 1985). *legA* and the pseudogene *legD* occur on a 13.5kbp fragment (Bown *et al.*,1985), *legE* on the 4.3kbp fragment (Rerie *et al.*,1990, Yaish, 1990) and *legB* and *legC* presumably on the other two fragments. The genes within this subfamily encode the L4 subunits (Matta *et al.*,1981) marked as *LegA-C* on fig. 23. cDNAs have been isolated which encode peptides sequenced from the L5 subunits (Matta *et al.*,1981) marked as *LegS* on fig. 23. The three subunits are arranged αN-αC-β on the cDNA (5'-3'), only the αN and β

subunits containing cysteine residues and remaining linked after proteolytic cleavage (Domoney *et al.*,1986b, Gatehouse,JA. & Gilroy,JS., unpublished work). There is some doubt as to the gene copy number of the *legS* subfamily although a maximum number of two seems likely (Domoney & Casey, 1985, Domoney *et al.*,1986a).

Although N-terminal sequences from the L1 (Matta *et al.*,1981) subunits, *legX* on fig. 23 have been determined (March *et al.*,1988), no cDNA or genes have yet been isolated from this subfamily so it is impossible to estimate its size. The remaining subunit pairs L2 and L3 (Matta *et al.*,1981), *legJ* and *legK* on fig. 23, comprise Casey's α^m subunits (Casey, 1979). Peptides sequenced from these match those of the *legJ* gene subfamily (Gatehouse *et al.*,1988), the subject of part of this work. From genomic blots it can be seen that this subfamily consists of three members (Domoney & Casey, 1985, Levasseur, 1988). It has been shown that *legK* is not expressed in FF and that the protein encoded by *legJ* is of similar Mr to that encoded by *legK*. The position of the *LegK* α-subunit on two-dimensional gel has been established in Birte as lying close to that of the L3 subunit or *legK* on fig. 23. The L3 subunit pair is therefore ascribed to *legJ* and the L2 subunit pair to a related gene, *legL*, postulated to have an enlarged acidic region at the C-terminus of the α-subunit.

CHAPTER TEN: ROOT PROTEIN RESULTS

Isolation of the major pea root protein

Root protein extraction

Total root protein was extracted from 1g of lyophilised roots from flowering FF pea plants and precipitated with ammonium sulphate to 90% saturation. The precipitate was resuspended in 1xPBS 20mM β-mercaptoethanol, this and the supernatant dialysed against the resuspension buffer and lyophilised. Samples of the ammonium sulphate precipitate and supernatants were electrophoresed on 17.5% acrylamide gel under reducing conditions (fig. 25a). No protein could be seen in the supernatant, the precipitate contained many bands of similar intensity predominantly in the 23-80k Mr range and a major band of 16,000 Mr.

Ammonium sulphate fractionation of root protein

To purify the major protein, ammonium sulphate cuts were performed on protein extracted from 10g of lyophilised roots. Ammonium sulphate was added successively to 45, 60, 75 and 90% saturation. The precipitates were resuspended in 1xPBS 20mM β-mercaptoethanol, dialysed against the resuspension buffer and lyophilised. Samples of these were electrophoresed on 17.5% polyacrylamide gel under reducing conditions. The gel, fig. 25b, shows that the majority of the 16k Mr protein is in the 60-75% ammonium sulphate fraction but this is heavily contaminated with other proteins.

Purification of the major protein by gel filtration

As the major protein band appeared to be reasonably well separated in size from contaminating proteins on polyacrylamide gel of the 60-75% ammonium sulphate fraction (fig. 25b), it was decided to purify this by

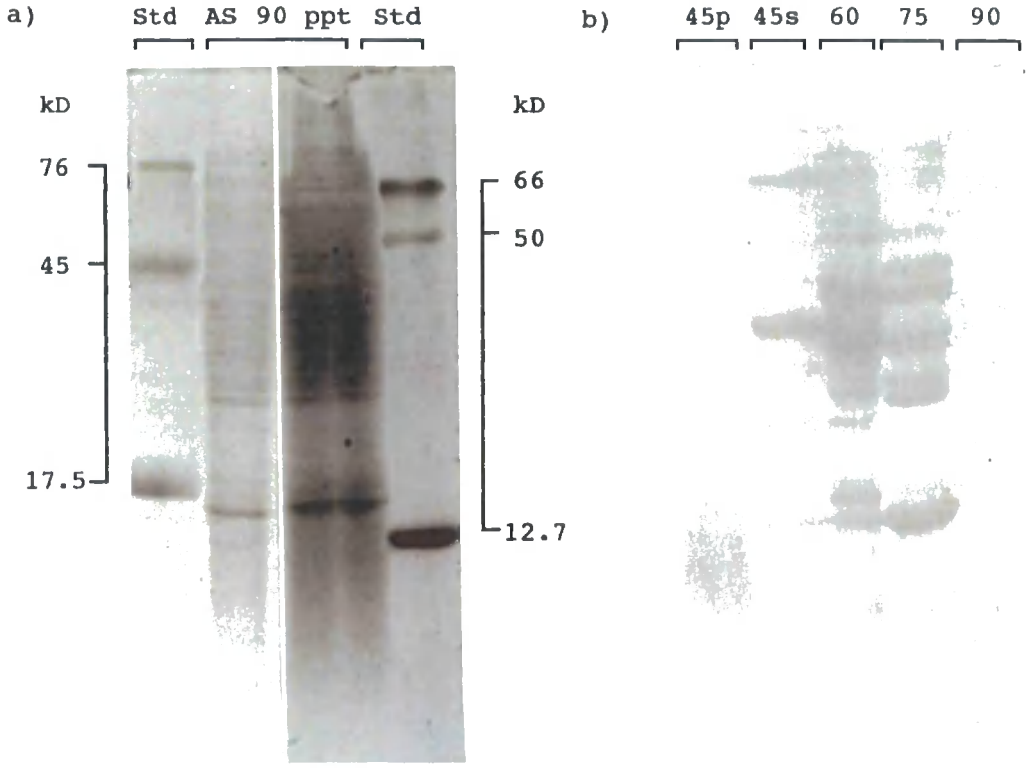
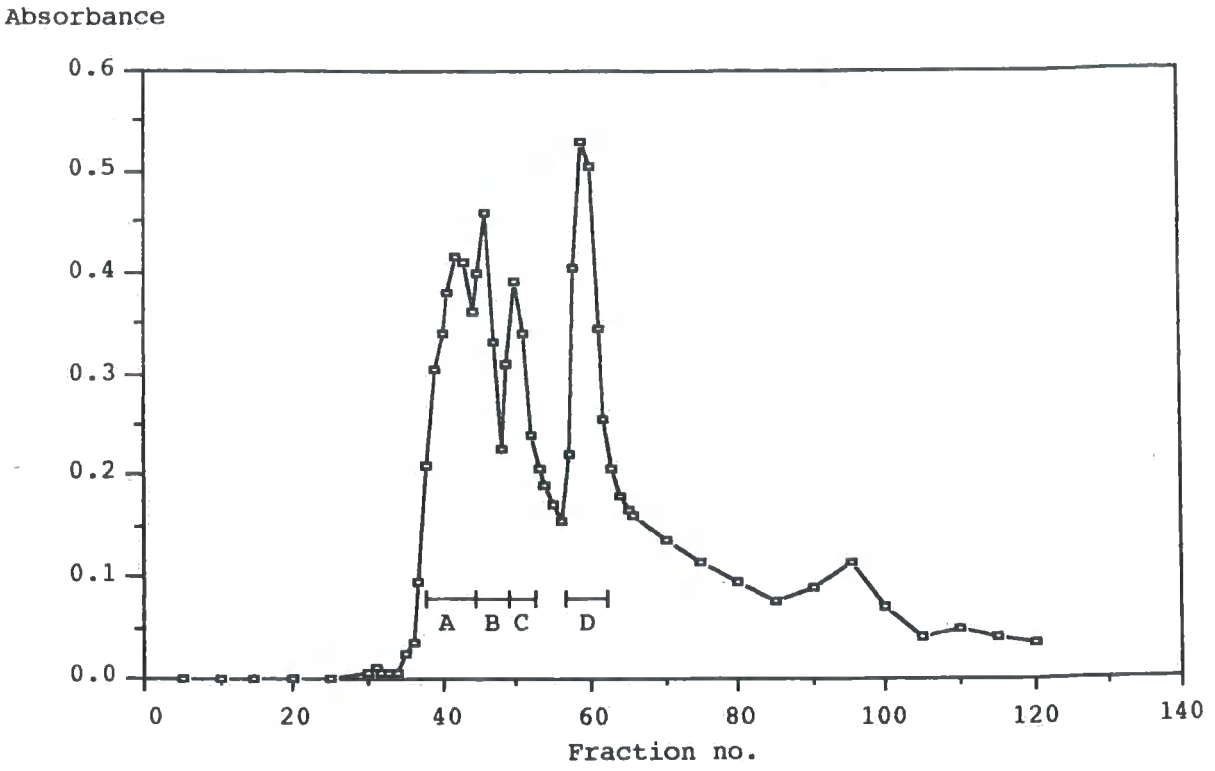


Fig. 25 Polyacrylamide gels of; a) 90% ammonium sulphate precipitable pea root protein, 100 and 500µg loaded, and b) ammonium sulphate fractions of pea root protein, 300µg loadings. Standard proteins were run on gel a) and their Mrs are indicated. Fractions on gel b) are the protein precipitable after the successive addition of ammonium sulphate to 45,60,75 and 90% saturation. The 45% ppt was resuspended, dialysed and the ppt after dialysis run separately (45p) from the supernatant (45s).



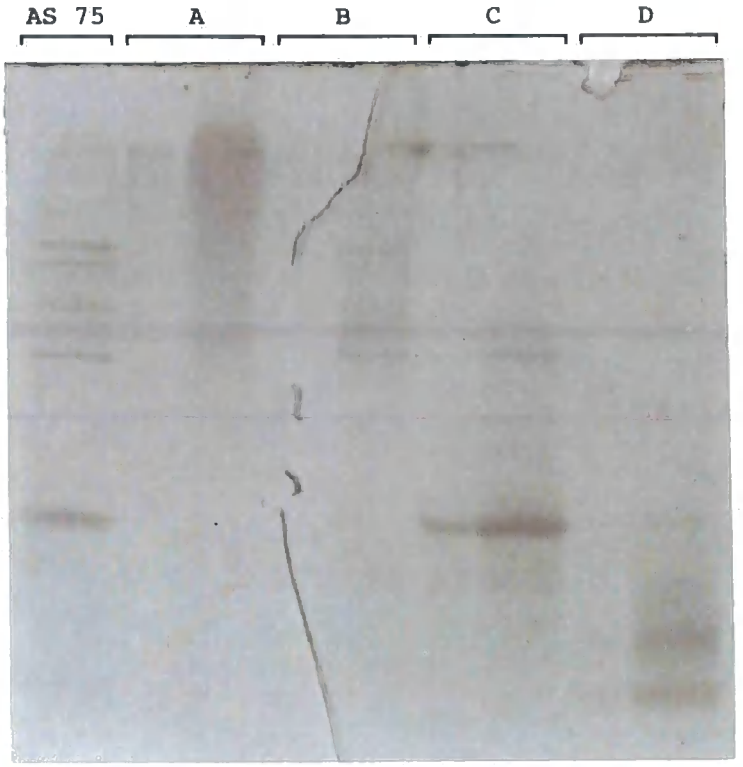


Fig. 26 Absorbance of fractions from root protein 60-75% ammonium sulphate fraction separated on gel filtration column (above) and polyacrylamide gel of the regions, A,B,C & D pooled from this (below). A 500µg sample of the starting material was run on gel (AS 75) and loadings of 200 and 500µg of each pool run.

gel filtration. To prevent aggregation of the protein, 20mg of this fraction was reduced and carboxymethylated. A 1cm diameter, 180cm column was packed with sephacryl S-200 in 70% formic acid and the lyophilised sample run through under gravity. Fractions were collected and the absorbance at 280nm monitored, peaks were pooled (fig. 26), diluted 15x with H_2O and lyophilised. Samples from the peaks were electrophoresed on 17.5% polyacrylamide gel under reducing conditions. The gel, fig. 26, shows that the third peak from the column contained pure (single major band on gel) 16k Mr protein.

N-terminal sequencing of the major root protein

Manual N-terminal sequencing was performed on the purified 16k root protein using DABITC (Hirano *et al.*,1982, Chang *et al.*,1978). The N-terminal sequence was G V F V F D/(W) D E Y V S T, with the sixth residue giving predominantly aspartate with some tryptophan. When the use of an automated sequencer (Applied Biosystems model 471A protein sequencer) became available, a similar sample was run on that, this gave the N-terminal sequence as G V F V F D D E Y V S T V A A P (P) K L Y K A, an identical result to that obtained manually.

Trial CNBr treatment of the major root protein

To determine whether the protein purified contained any internal methionine residues, an attempt was made to cleave it with CNBr. 0.2mg of the gel filtration column fraction were treated with CNBr, incubating for 24hr and run on polyacrylamide gel. No difference could be seen between CNBr treated and non-treated fraction (data not shown).

Sequencing of tryptic peptides

As CNBr had failed to cleave the 16k Mr root protein, no internal sequence data could be obtained from its products. The gel filtration column peak was digested with trypsin and the products passed through HPLC, peaks were collected and sequenced using DATBITC (as described in; Gatehouse *et al.*,1982a & Chang *et al.*,1978). Unfortunately, due to an electrical fault, the samples were destroyed through over heating after a few cycles. Results obtained were:

D A D	G D A
E A Q	A L/I E G Y
G V F	F V Q

Trial oligonucleotide hybridisation to root RNA

Using the N-terminal sequence, an oligonucleotide sequence with the least redundancy was deduced:

AA seq	G	V	F	V	F	D	D	E	Y	V	S	T	V	A	P	P	K	L	Y	K	A
DNA			C		C	C	C	A	C								AC		C	A	
coding	GGNGTNTT			GTNTT		GA	GA	GA	TA	GTNTCNACNGTNGCNCCNCCNAA							TNTA		AA	GCN	
seq			T		T	T	T	G	T								GT		T	G	
least redundent	----------																				
					G	G	G	T	G												
compliment					AA	CT	CT	CT	AT	CA											
					A	A	A	C	A												

The sequence, A C G/A T A T/C T C G/A T C G/A T C G/A A A, the *reverse* compliment of the least redundant sequence was synthesised.

To check whether this could be used as a probe against a cDNA library, a northern blot was performed. RNA was isolated from roots of flowering FF pea plants, run on agarose gel and blotted. The filter was probed with the end labelled oligonucleotide, pre-hybridising and hybridising at 40°C and washing to 1xSSC 1%SDS at 40°C for 1hr. No hybridisation could be seen to the filter (data not shown), although

experiments under similar conditions had been successful with convicilin oligonucleotide probes (qv.). The hybridisation was repeated at 25°C, washing at this temperature, initially to 5xSSC and then 3xSSC, but again no hybridisation could be detected above background.

CHAPTER ELEVEN: ROOT PROTEIN DISCUSSSION

Purification of the major root protein

The purification of the 16k Mr protein, which was the most abundant species present in roots from flowering FF pea plants, proved to be relatively straightforward. A 60-75% ammonium sulphate cut followed by gel filtration yielded a single band on polyacrylamide gel. To prevent oxidation β-mercaptoethanol was used and to prevent proteolysis leupeptin and PMSF were added in the initial extraction. There were signs that the protein was aggregating to higher Mr forms after extraction and so to prevent disulphide bond formation, it was reduced and carboxymethylated.

Identification of the major root protein

The N-terminal sequence of the isolated protein was used to search the NBRF protein database, but no sequence with significant homology could be found. Due to the small amount of work so far undertaken on root proteins, this is not altogether surprising, despite the abundant nature of this protein. Pea lectin has been isolated from roots and cotyledons (Gatehouse & Boulter,1980) and contains a subunit of a similar Mr (17K) to the root protein reported here. However, the gene sequence of pea lectin (Gatehouse *et al.*,1987) encodes a protein with no homology to the peptides obtained from this protein nor could any homology be found to other seed proteins. This protein bears no similarity to the hydroxyproline-rich structural proteins which are present in quantity in root cell walls (Showalter & Varner, 1989). The root expressed enzymes studied are presumably present at lower levels than this protein.

Studies have been undertaken on protein from specialised underground plant organs such as tubers and legume root nodules. In

potato tubers, patatin accounts for up to 40% of the soluble protein (Edwards & Coruzzi, 1990) and although tubers are derived from stem tissue, expression of patatin encoding genes has been detected in root tissue although at a substantially (100x) lower level (Pikaard *et al.*, 1987). Again the cDNA encoded protein (Mignery *et al.*,1984) bears no homology to that isolated here from pea roots and the same is true for the protease inhibitors isolated from potato (Richardson, 1991). Although most genes regarded as nodule specific have now been found to be expressed in other tissues (Edwards & Coruzzi, 1990) this is at a low level. No evidence of nodulation could be seen in the plants grown hydroponically for this work.

Attempt to isolate a cDNA encoding the major root protein

As only the N-terminus of the root protein had been successfully sequenced for a reasonable distance, an oligonucleotide was prepared against the least redundant DNA sequence encoding residues from this. A 17mer with 32 fold redundancy was prepared, labelled and tested against root RNA on a N blot. Although the conditions used had been successful with convicilin oligos and the initial temperature of washing (40°C) was below that calculated for the melting temperature of the oligo containing the most (A+T), 42°C (Maniatis *et al.*,1982), no hybridisation above background could be seen. Hybridisation and washing at lower temperature also failed to give a positive result.

As the detection of mRNA had failed with this oligo, it was decided not to proceed to a cDNA library screen especially as the 5' end of cDNA encoding the N-terminus of proteins is that most likely to be lost during synthesis. Had more time been available it would have been possible to purify more of this protein for sequencing of tryptic peptides. Further oligos could then have been constructed from internal

protein sequence and these used either to screen mRNA and cDNA directly or, using PCR, to amplify probes for screening.

Plasmid cDNA Library

A cDNA library was constructed in the vector pUC18 using poly(A) enriched RNA from 5 daf PP pea pods. DNA was prepared from twelve random recombinants, and the insert size determined. Inserts ranged in size from 200 to 1500bp, although most inserts were in the range 200-500bp. 1152 recombinants were preserved in 40% glycerol in twelve duplicate 96 well microtitre plates, and were designated pPP1-pPP1152.

Trial Screen of plasmid cDNA library

Using a 48 pronged fork, 4 replicate sets of colonies from the preserved cDNA library were grown and amplified on nitrocellulose filters. Colonies were lysed and the DNA bound to the nitrocellulose by baking.

To test whether genes from the anthocyanin biosynthesis pathway could be isolated from this library, one set of filters was probed with the insert from *Phaseolus vulgaris* chalcone synthase cDNA, pCHS1 (Ryder *et al.*,1987), and another set with a fragment from the maize *A1* gene, pMu0.6Pst from c10Mu (O'Reilly *et al.*,1985). After washing the filters to 2xSSC, 2x30mins at 65°C, four recombinants remained hybridised to by the pCHS1 probe but no convincing hybridisation could be observed on the filters probed with *A1*.

DNA was isolated from the pCHS1 positive clones, restricted, electrophoresed on agarose gel, blotted onto nitrocellulose filter and probed with the probe used above, washing to 1xSSC, 2x30mins at 65°C. DNA inserts from three of the clones still hybridised to the probe under these conditions. The DNA sequences of the three inserts were determined manually and the results are presented in fig. 27. pPP166 is identical to pPP888, having an insert 193bp in length containing an ORF of 117bp

Fig. 27 Nucleotide sequences for pea pod chalcone synthase cDNAs pPP166 and pPP372. The derived amino acid sequence from pPP166 is shown below the DNA sequence and those residues differing in the derived sequence from pPP372 are shown above. The consensus sequence for this region from *Phaseolus vulgaris* chalcone synthase cDNA (CHS) is also presented. A putative polyadenylation signal in the pPP166 sequence is underlined.

encoding 39 aa pPP372 has an insert 294bp in length with an ORF of 102bp (34 aa). The ORFs were confirmed by homology to the determined aa sequence from *Phaseolus vulgaris* CHS (Ryder *et al.*,1987). In the 102bp region of overlap in the coding sequence of the two pea CHS cDNA clones there are 14 base differences resulting in eight aa substitutions. The sequence 3' from the stop codon is conserved between the two for 23bp before they diverge.

Differential screen of plasmid cDNA library

The remaining two replicate sets of filters containing the cDNA library were independently probed with labelled cDNA prepared from 5 daf pea pods, from either the purple podded line PP or the standard (green podded) line FF. Hybridisation was allowed to proceed for 60 hrs and washing was to 2xSSC, 2x30mins, all at 65°C. DNA was prepared from 37 clones which appeared to be hybridised to more strongly by the cDNA prepared from the purple podded line than the green podded line. The insert sizes from these when determined by restriction and agarose gel electrophoresis were in the range 100-1000bp, 30 of which were below 300bp in length.

On the basis of size and degree of differential hybridisation, inserts from 13 of these cDNAs were isolated, labelled and used to probe northern blots containing total RNA from FF, PP and mutant green pods from the PP line (GP). Washing was to 0.1xSSC, 0.1%SDS at 50°C, for 2x20 mins. DNA from eight inserts hybridised non-specifically to ribosomal RNA in all lines, two hybridised with equal intensity to single mRNAs from all three pea lines, and three hybridised to single mRNAs more strongly in the PP lines than the FF lines. These results are presented in fig. 28.

Differentially expressed plasmid cDNAs

Clone pPP590 hybridised to a 1.75-1.8kb mRNA, strongly in PP and weakly in GP and FF RNAs, with the intensity of the band in GP 16% of PP and in FF 21% of PP (mean percentages of peak areas from autoradiographs of duplicate gels scanned by laser densitometer). pPP812 hybridised to a 500b mRNA, strongly in PP, weakly in GP (with the intensity of the GP band 16% of the PP band) and no hybridisation could be detected to FF RNA. pPP927 hybridised to a 2.2kb mRNA, strongly in PP and GP RNAs and weakly in FF, with the intensity of the FF band 5% of that in the PP track (GP when scanned gave 81% of PP but this figure is low due to irregularities in the band see fig. 28).

Manual sequencing was performed on these cDNAs. pPP590 (fig. 32) contains an insert of 374bp with one continuous ORF predicting a proline-rich aa sequence. pPP812 (fig. 29) contains a 274bp insert ORFs present in this sequence did not show significant homology to any polypeptide sequences in the NBRF protein database. A consensus polyadenylation signal occurs in a region 50bp from one end of the pPP812 sequence. pPP927 (fig. 37) contains a 241bp insert. One plausible ORF of 37 aa is present, and this is followed by a consensus

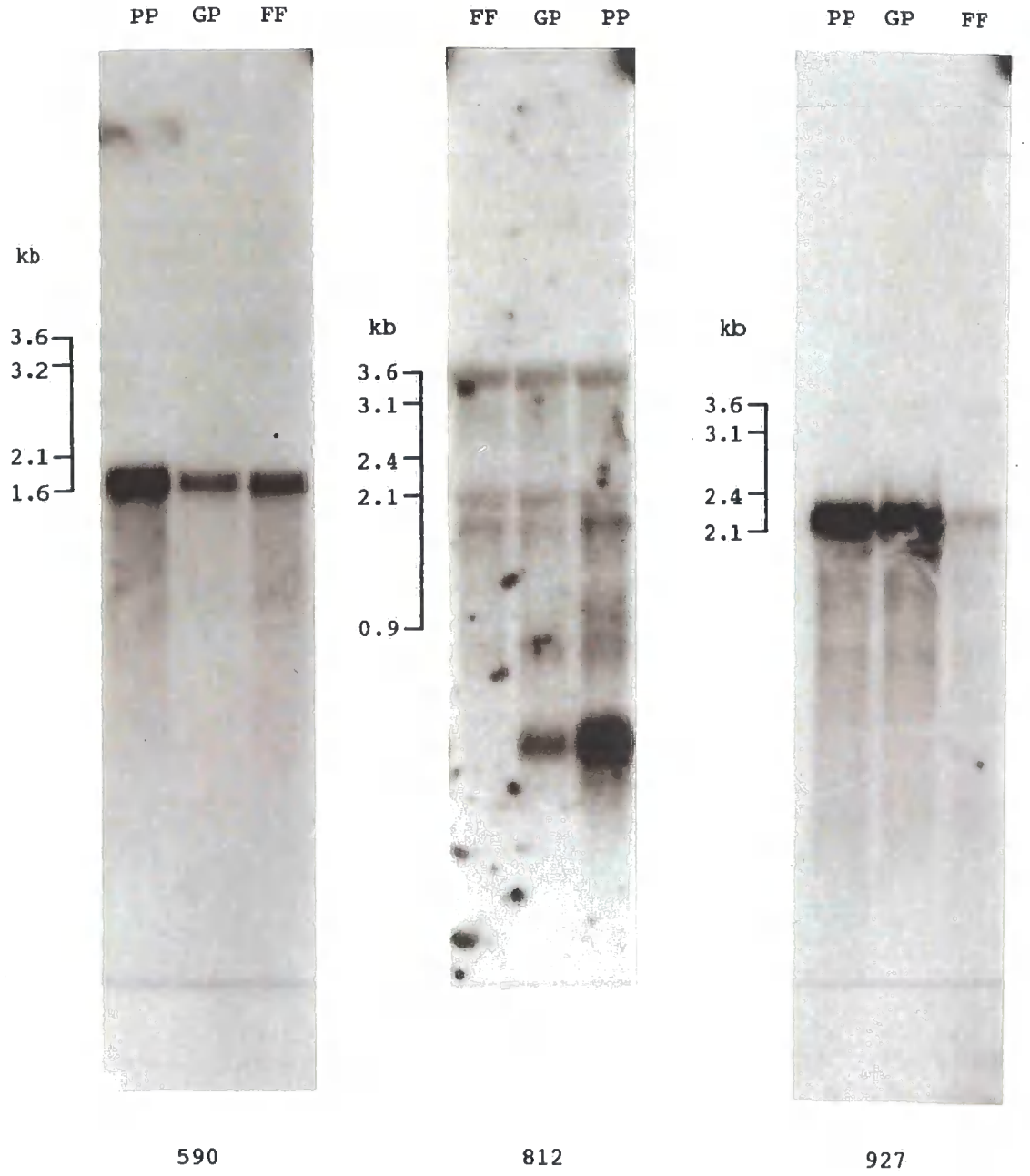


Fig. 28 Northern blots of total RNA from PP, GP, and FF pods probed with labelled pod cDNAs pPP590, pPP812 and pPP927. The position standard RNAs migrated to on the original gels is indicated for each filter.

Sequenced strand

```
GGAATACTCATCTCTGTTTTTATTGTTTTTCTTGTTTCTTCAACCGTGGAAGGTTCTACTTCCGGTTCCGGGGGTGGAGTCACCCTTCCTGAATCATGCACTCCACCGAAGTGTGAGTGC
 G  I  L  I  S  V  F  I  V  F  L  V  S  S  T  V  E  G  S  T  S  G  S  G  G  G  V  T  L  P  E  S  C  T  P  P  K  C  E  C
  E  Y  S  S  L  F  L  L  F  F  L  F  L  Q  P  W  K  V  L  L  P  V  P  G  V  E  S  P  F  L  N  H  A  L  H  R  S  V  S  A
   N  T  H  L  C  F  Y  C  F  S  C  F  F  N  R  G  R  F  Y  F  R  F  R  G  W  S  H  P  S  *  I  M  H  S  T  E  V  *  V  Q

AAAGTTCCACCTTGCTTTCCCCGTGCCGGCGCCGCCTCCACCACCTCCGCCGCTCTTCCACCACAGAATGATAATTAGCATTATCATTTCTATGTAATAAATAAAAATATCTGTATCTCT
 K  V  P  P  C  F  P  R  A  G  A  A  S  T  T  S  A  A  L  P  P  Q  N  D  N  *  H  Y  H  F  Y  V  I  N  K  N  I  C  I  S
  K  F  H  L  A  F  P  V  P  A  P  P  P  P  P  P  P  L  F  H  H  R  M  I  I  S  I  I  I  S  M  *  *  I  K  I  S  V  S  L
   S  S  T  L  L  S  P  C  R  R  R  L  H  H  L  R  R  S  S  T  T  E  *  *  L  A  L  S  F  L  C  N  K  *  K  Y  L  Y  L  Y

ATGTTGTACCTATAATGAGTACAACCTCGTGCCG
 M  L  Y  L  *  *  V  Q  P  R  A
  C  C  T  Y  N  E  Y  N  L  V  P
   V  V  P  I  M  S  T  T  S  C
```

Complementary strand

```
CGGCACGAGGTTGTACTCATTATAGGTACAACATAGAGATACAGATATTTTTATTTATTACATAGAAATGATAATGCTAATTATCATTCTGTGGTGGAAGAGCGGCGGAGGTGGTGGAGG
 R  H  E  V  V  L  I  I  G  T  T  *  R  Y  R  Y  F  Y  L  L  H  R  N  D  N  A  N  Y  H  S  V  V  E  E  R  R  R  W  W  R
  G  T  R  L  Y  S  L  *  V  Q  H  R  D  T  D  I  F  I  Y  Y  I  E  M  I  M  L  I  I  I  L  W  W  K  S  G  G  G  G  G  G
   A  R  G  C  T  H  Y  R  Y  N  I  E  I  Q  I  F  L  F  I  T  *  K  *  *  C  *  L  S  F  C  G  G  R  A  A  E  V  V  E  A

CGGCGCCGGCACGGGGAAAGCAAGGTGGAACTTTGCACTCACACTTCGGTGGAGTGCATGATTCAGGAAGGGTGACTCCACCCCCGGAACCGGAAGTAGAACCTTCCACGGTTGAAGAAA
 R  R  R  H  G  E  S  K  V  E  L  C  T  H  T  S  V  E  C  M  I  Q  E  G  *  L  H  P  R  N  R  K  *  N  L  P  R  L  K  K
  G  A  G  T  G  K  A  R  W  N  F  A  L  T  L  R  W  S  A  *  F  R  K  G  D  S  T  P  G  T  G  S  R  T  F  H  G  *  R  N
   A  P  A  R  G  K  Q  G  G  T  L  H  S  H  F  G  G  V  H  D  S  G  R  V  T  P  P  P  E  P  E  V  E  P  S  T  V  E  E  T

CAAGAAAAACAATAAAAACAGAGATGAGTATTCC
 Q  E  K  Q  *  K  Q  R  *  V  F
  K  K  N  N  K  N  R  D  E  Y  S
   R  K  T  I  K  T  E  M  S  I
```

Fig. 29 Nucleotide and derived amino acid sequences from pPP812 cDNA. Consensus polyadenylation signals are underlined.

polyadenylation signal in a region 100bp 3' from the possible stop codon. On searching the NBRF protein database this ORF was found to show homology to the aa sequence encoded by a pectinesterase cDNA from tomato (Ray *et al.*,1988).

Tissue specificity of differentially expressed cDNAs

Samples of pod, leaf and petal total RNA from the PP line were run on gel, blotted and probed with a mixture of labelled pPP590, pPP812 and pPP927 cDNAs, washing to 0.1xSSC 0.1%SDS at 50°C. The result (fig. 30a) is that all three cDNAs hybridise strongly to their source tissue - pod, but not to leaf or petal RNA. The amount of petal RNA run on this gel was in fact less than the other tissues (4.3μg compared with 10μg of pod and leaf). A similar experiment was performed using equal amounts of RNA (fig. 30b) and probing with only pPP590 and pPP812, washing to the same stringency. Again both cDNAs hybridised strongly to pod RNA and not to leaf RNA, but weak hybridisation could be seen, by pPP590 only, to petal RNA.

Phage lambda cDNA library

As the previous cDNA library had produced predominantly very short cDNA, a second library was constructed, using mRNA from the same 5 daf PP pea pods, but this time, using the λgt10 vector. A library containing $6x10^5$ distinct cDNA transformants was obtained. Duplicate nitrocellulose filters each containing $2.7x10^4$ plaques were screened with a mixture of labelled inserts from the three cDNAs isolated from the plasmid cDNA library; pPP590, pPP812 and pPP927. Plugs containing positively hybridising plaques were removed and phage from them plated and screened with the three cDNAs separately. All plugs contained pPP927 positive plaques but none were hybridised to by pPP590 or pPP812.

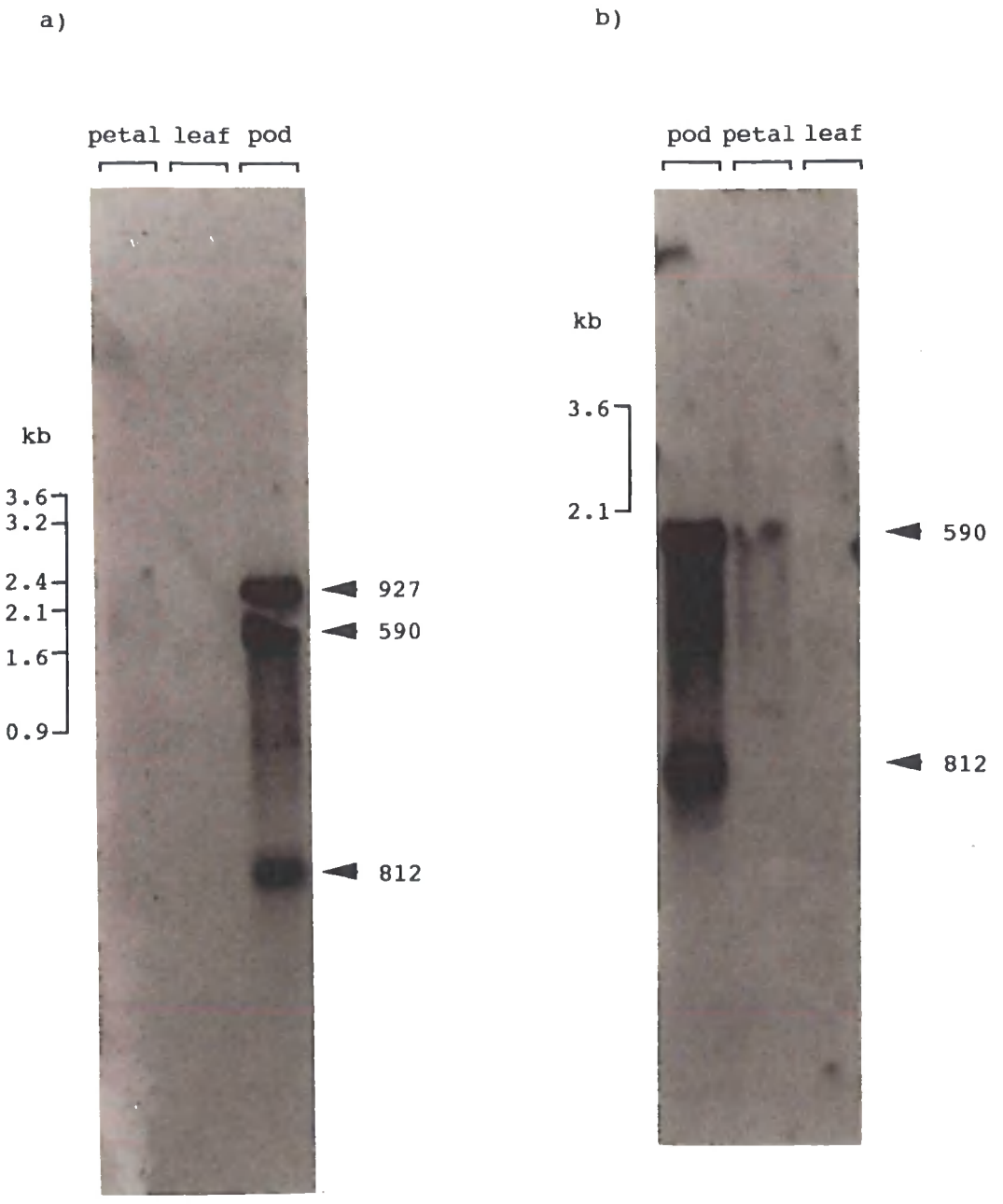


Fig. 30 Northern blots of total RNA from various tissues from line PP probed with mixed pod cDNAs. Blot a) is probed with pPP590, pPP812 and pPP927, blot b) is probed with pPP590 and pPP812. The position standard RNAs ran to on the original gels is indicated. The identity of the hydridising species (judged from the sizes determined for the separate cDNAs, fig. 28) is shown for each filter.

pPP927 positive lambda cDNA

Four of the pPP927 positive plaques were purified, DNA prepared from them, restricted to excise the insert and electrophoresed on agarose gel. The gel was blotted and the nitrocellulose filter probed with pPP927 insert. The λ cDNAs contained inserts hybridised to by pPP927 in the range 400-600bp. The insert from one of the larger clones was amplified by PCR and cloned into plasmid vector. The DNA sequence of the insert from this plasmid, designated pPPL927, was determined by automated sequencing and is shown on fig. 37.

This cDNA is 478bp in length. It includes all of the plasmid cDNA sequence and corresponding regions are identical. The λ cDNA sequence extends an extra 69bp 3', 1t contains two further consensus polyadenylation signals, and at the 3' end, a 15bp poly(A) tail. It also extends an additional 168bp in the 5' direction extending the ORF by 56 aa; the additional aa sequence continues the homology with tomato pectinesterase. It is interesting to note that although this insert was amplified by PCR, no differences were observed between the two sequences over the 241bp of the plasmid cDNA. *Taq* polymerase errors should be anticipated in the sequence from these products (Karlovsky, 1990).

Genomic library construction

As the cDNA libraries had failed to produce full length cDNA, it was decided to proceed to a genomic library using PP genomic DNA and the vector λGEM-11. A library containing $9x10^5$ distinct transformants was obtained. Duplicate sets of nitrocellulose filters containing $4x10^6$ plaques were screened in turn with labelled inserts from clones isolated from the λ cDNA library; pPPL927 and the plasmid cDNA library; pPP590 and pPP812. Plugs containing duplicate positively hybridising regions were taken and plaques purified from the pPP590 and pPPL927 screens. No

plaques could be purified from the apparently positive plugs taken after screening with pPP812.

pPP590 positive genomic clone

Six clones hybridised to by pPP590 were purified to single plaques, DNA was prepared from these, restricted and electrophoresed on agarose gels. All six clones were identical. Using the results from the gels and those obtained from probing blots of these gels with labelled pPP590 insert, a restriction map for this genomic clone, designated PP590, was prepared (fig. 31). The region cloned is 15.2kbp in length. A 1kbp *Hind* III fragment 5.2kbp from the left λ arm was hybridised to by the cDNA and this and adjacent regions were subcloned into a plasmid vector. The plasmid subclones were further restriction mapped, M13 subclones prepared and automated DNA sequencing performed on these. A detailed restriction map and sequencing strategy is also shown on fig. 31. 3234bp were sequenced, in both directions with all restriction sites sequenced through.

PP590 sequence

The longest ORF occurring in the sequence obtained corresponding to the ORF with no stop codons in the cDNA (fig. 32) is 1056bp in length, encoding 352 aa. There is no evidence for the presence of any intron sequences interupting the ORF. The start is within the *Hind* III fragment hybridised to by the cDNA and the 3' end of the ORF extends 95bp into the adjacent 1.3kbp *Hind* III to *Eco* RI fragment. The nucleotide sequence is identical to that of the cDNA clone suggesting that the cDNA is a product of this gene.

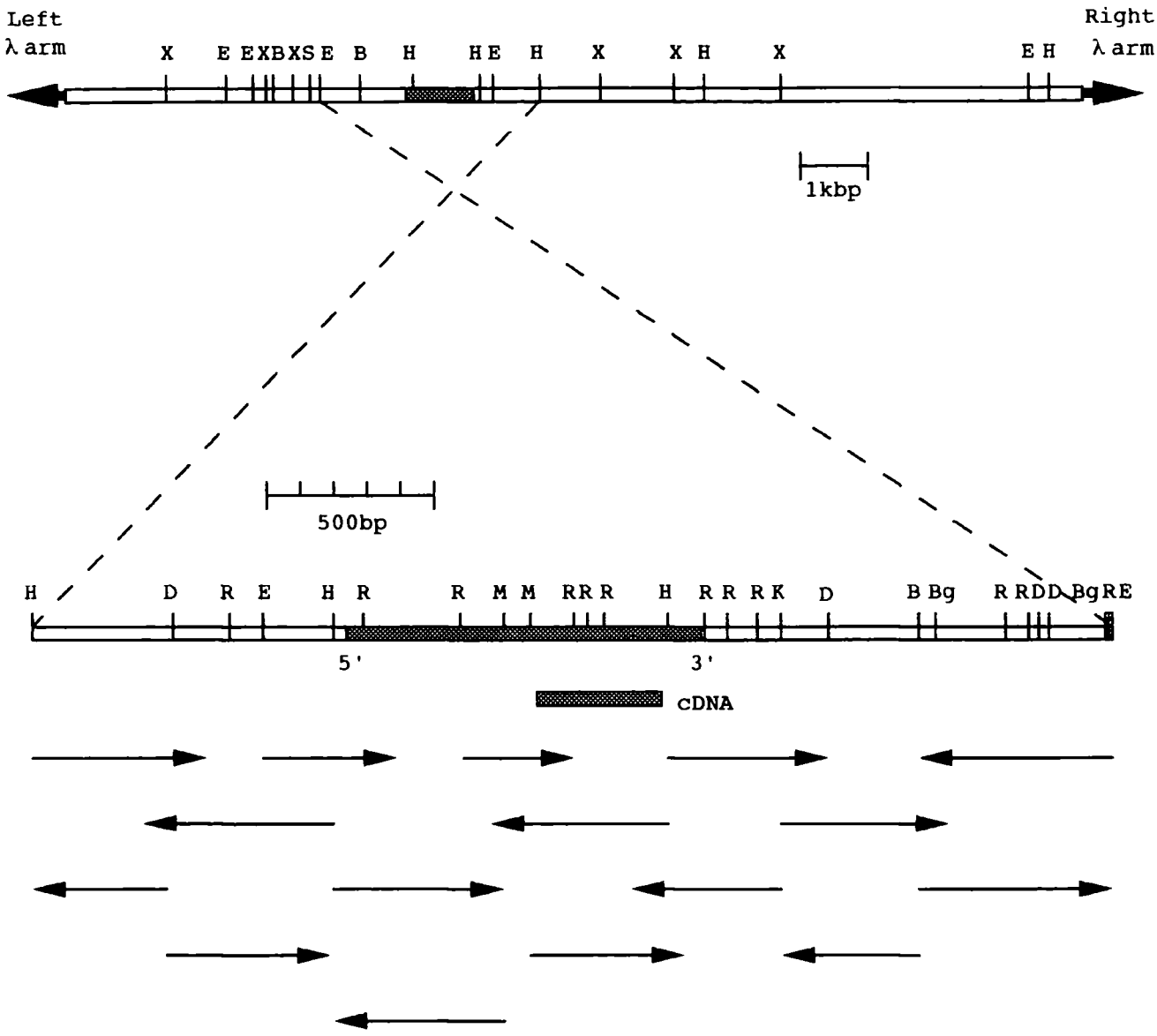
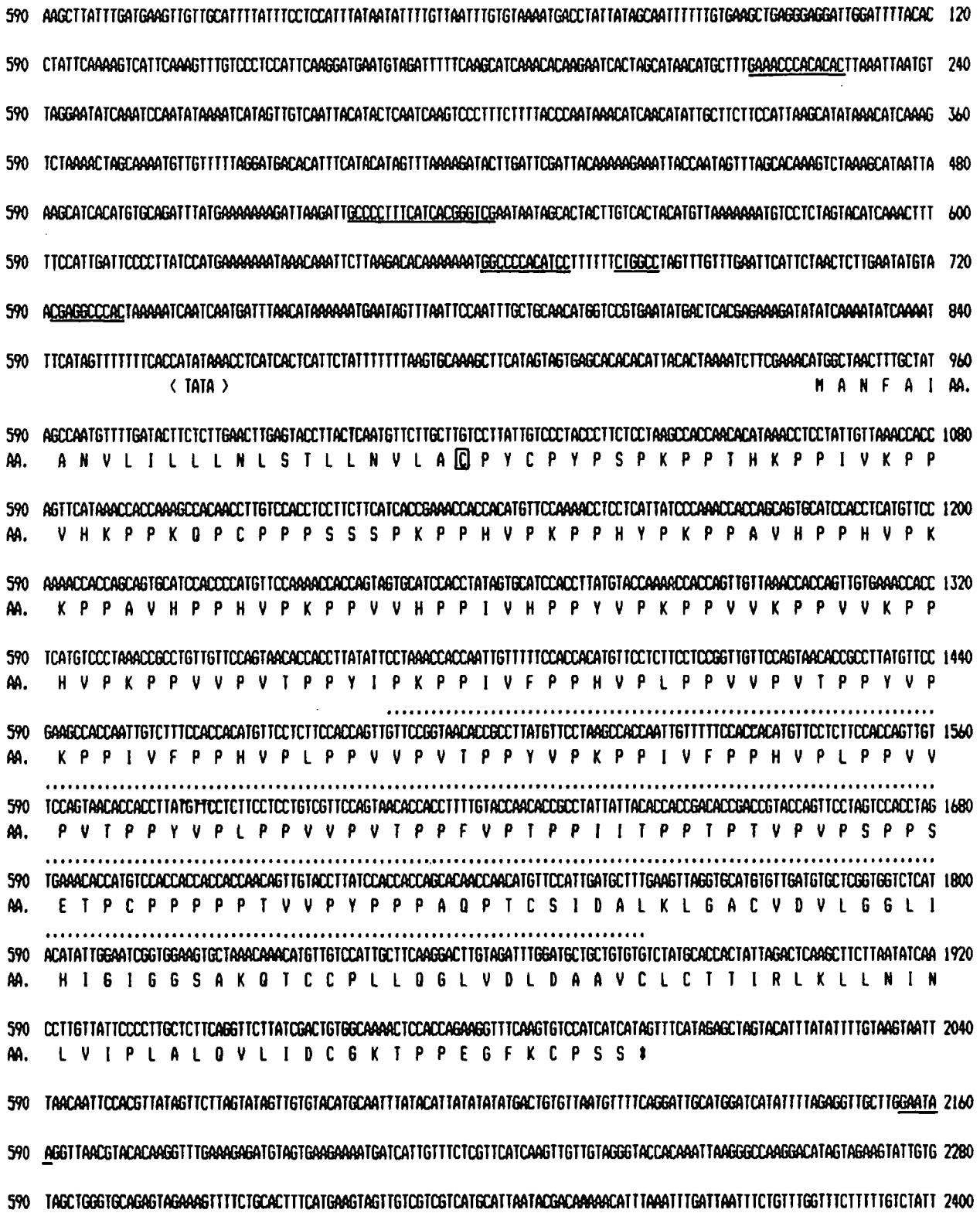


Fig. 31 Restriction maps of PP590 genomic clone and the sequenced region from it. Restriction enzymes are abbreviated as: B, *Bam* HI; Bg, *Bgl* II; D, *Dra* I; E, *Eco* RI; H, *Hind* III; K, *Kpn* I; M, *Msp* I; R, *Rsa* I; X, *Xba* I. The coding region is highlighted, the position of the cDNA shown and arrows below the sequenced region represent individual sequencing runs.


590 AAGCTTATTTGATGAAGTTGTTGCATTTTATTTCCTCCATTTATAATATTTTGTTAATTTGTGTAAAATGACCTATTATAGCAATTTTTTGTGAAGCTGAGGGAGGATTGGATTTTACAC 120
590 CTATTCAAAAGTCATTCAAAGTTTGTCCCTCCATTCAAGGATGAATGTAGATTTTTCAAGCATCAAACACAAGAATCACTAGCATAACATGCTTTGAAACCCACACACTTAAATTAATGT 240
590 TAGGAATATCAAATCCAATATAAAATCATAGTTGTCAATTACATACTCAATCAAGTCCCTTTCTTTTACCCAATAAACATCAACATATTGCTTCTTCCATTAAGCATATAAACATCAAAG 360
590 TCTAAAACTAGCAAAATGTTGTTTTTAGGATGACACATTTCATACATAGTTTAAAAGATACTTGATTCGATTACAAAAAGAAATTACCAATAGTTTAGCACAAAGTCTAAAGCATAATTA 480
590 AAGCATCACATGTGCAGATTTATGAAAAAAAGATTAAGATTGCCCCTTTCATCACGGGTCGAATAATAGCACTACTTGTCACTACATGTTAAAAAAATGTCCTCTAGTACATCAAACTTT 600
590 TTCCATTGATTCCCCTTATCCATGAAAAAAATAAACAAATTCTTAAGACACAAAAAAATGGCCCCACATCCTTTTTTCTGGCCTAGTTTGTTTGAATTCATTCTAACTCTTGAATATGTA 720
590 ACGAGGCCCACTAAAAATCAATCAATGATTTAACATAAAAAATGAATAGTTTAATTCCAATTTGCTGCAACATGGTCCGTGAATATGACTCACGAGAAAGATATATCAAAATATCAAAAT 840
590 TTCATAGTTTTTTTCACCATATAAACCTCATCACTCATTCTATTTTTTTAAGTGCAAAGCTTCATAGTAGTGAGCACACACATTACACTAAAATCTTCGAAACATGGCTAACTTTGCTAT 960
(TATA) M A N F A I AA.
590 AGCCAATGTTTTGATACTTCTCTTGAACTTGAGTACCTTACTCAATGTTCTTGCTTGTCCTTATTGTCCCTACCCTTCTCCTAAGCCACCAACACATAAACCTCCTATTGTTAAACCACC 1080
AA. A N V L I L L L N L S T L L N V L A C P Y C P Y P S P K P P T H K P P I V K P P
590 AGTTCATAAACCACCAAAGCCACAACCTTGTCCACCTCCTTCTTCATCACCGAAACCACCACATGTTCCAAAACCTCCTCATTATCCCAAACCACCAGCAGTGCATCCACCTCATGTTCC 1200
AA. V H K P P K Q P C P P P S S S P K P P H V P K P P H Y P K P P A V H P P H V P K
590 AAAACCACCAGCAGTGCATCCACCCCATGTTCCAAAACCACCAGTAGTGCATCCACCTATAGTGCATCCACCTTATGTACCAAAACCACCAGTTGTTAAACCACCAGTTGTGAAACCACC 1320
AA. K P P A V H P P H V P K P P V V H P P I V H P P Y V P K P P V V K P P V V K P P
590 TCATGTCCCTAAACCGCCTGTTGTTCCAGTAACACCACCTTATATTCCTAAACCACCAATTGTTTTTCCACCACATGTTCCTCTTCCTCCGGTTGTTCCAGTAACACCGCCTTATGTTCC 1440
AA. H V P K P P V V P V T P P Y I P K P P I V F P P H V P L P P V V P V T P P Y V P
590 GAAGCCACCAATTGTCTTTCCACCACATGTTCCTCTTCCACCAGTTGTTCCGGTAACACCGCCTTATGTTCCTAAGCCACCAATTGTTTTTCCACCACATGTTCCTCTTCCACCAGTTGT 1560
AA. K P P I V F P P H V P L P P V V P V T P P Y V P K P P I V F P P H V P L P P V V
590 TCCAGTAACACCACCTTATGTTCCTCTTCCTCCTGTCGTTCCAGTAACACCACCTTTTGTACCAACACCGCCTATTATTACACCACCGACACCGACCGTACCAGTTCCTAGTCCACCTAG 1680
AA. P V T P P Y V P L P P V V P V T P P F V P T P P I I T P P T P T V P V P S P P S
590 TGAAACACCATGTCCACCACCACCACCAACAGTTGTACCTTATCCACCACCAGCACAACCAACATGTTCCATTGATGCTTTGAAGTTAGGTGCATGTGTTGATGTGCTCGGTGGTCTCAT 1800
AA. E T P C P P P P P T V V P Y P P P A Q P T C S I D A L K L G A C V D V L G G L I
590 ACATATTGGAATCGGTGGAAGTGCTAAACAAACATGTTGTCCATTGCTTCAAGGACTTGTAGATTTGGATGCTGCTGTGTGTCTATGCACCACTATTAGACTCAAGCTTCTTAATATCAA 1920
AA. H I G I G G S A K Q T C C P L L Q G L V D L D A A V C L C T T I R L K L L N I N
590 CCTTGTTATTCCCCTTGCTCTTCAGGTTCTTATCGACTGTGGCAAAACTCCACCAGAAGGTTTCAAGTGTCCATCATCATAGTTTCATAGAGCTAGTACATTTATATTTTGTAAGTAATT 2040
AA. L V I P L A L Q V L I D C G K T P P E G F K C P S S *
590 TAACAATTCCACGTTATAGTTCTTAGTATAGTTGTGTACATGCAATTTATACATTATATATATGACTGTGTTAATGTTTTCAGGATTGCATGGATCATATTTTAGAGGTTGCTTGGAATA 2160
590 AGGTTAACGTACACAAGGTTTGAAAGAGATGTAGTGAAGAAAATGATCATTGTTTCTCGTTCATCAAGTTGTTGTAGGGTACCACAAATTAAGGGCCAAGGACATAGTAGAAGTATTGTG 2280
590 TAGCTGGGTGCAGAGTAGAAAGTTTTCTGCACTTTCATGAAGTAGTTGTCGTCGTCATGCATTAATACGACAAAAACATTTAAATTTGATTAATTTCTGTTTGGTTTCTTTTTGTCTATT 2400

590 TGTTTGATTGCGTGTGATGCTCTGTGTTTGCATGTCAGCGTTTGTAACTTCACCCGTGAATGCATTGGAAATGTATAAGTTCTTTTTTTTTATGTATTTTCTTAAAAAAATGATTATCAT 2520

590 GCTACACGTTGTTTCTTTCTTTTTTTGGTCAAGACACGTTTCTTTAATAATTATGAATATGTAGAGTATATATTGGACCAGGGTGGAAGTGATAGCATGATTGCATGTGATTTTGAACAA 2640

590 TAAAATTGGATCCATATATCAACTTAAAAAGTTAAATCACTAGAACAAATAATGATATATAGATCTTTCGGTTCTTATTATGATATATATAATAATAAATGCTTATTTTGAATCTTCTAT 2760

590 TGTTGACGGATAGTTGTAGTCATTCATTCTATTGCAAATTTTCAATACATAATATTATTATGAGTCGGAGACAAAACTTGTCTTCAGTGATTTTGTTTGGTGAGCATATCGTGTTTTTTT 2880

590 CTTTGATCTTCCATAATTTACGTTAAGTGTACGTGCATGTTCTATTAATTAATATTCTAAATCATATAGTAATAATAATTGGCTTATATTAGTGCAAAGCGTACAAAATTGAAGAGATTA 3000

590 GACTCGTGAGTTCCTTTAAAAAAACAAAAAGATATTGTTTTAAAAATAAAAATAAAATCGTGTCTTGTCTTATATTAGAAAAACAAAAACCACACATTTTTATAATTATATATTAAAATT 3120

590 ATTAATATATGAGCATCCACAACTTCTTTACAACAGCCTAGCAAGCAAATATCTTAGTATTGACTCAGCTTTATTGGTAATTGTGCATGACATTAGATCTAAGGGTACGAATTC 3234

Fig. 32 Nucleotide (590) and derived amino acid sequence (A.A.) of PP590 (*ptxA*) gene. The extent of the plasmid cDNA, pPP590 is indicated thus ..... G/C rich regions in the 5' flanking sequence and consensus polyadenylation signals are underlined. The N-terminal residue predicted for PP590 is boxed.

The proline-rich sequence

On inspection of the derived aa sequence from PP590 it is apparent that it contains a major proline rich region. The first proline of this region is at residue 26, and it extends to residue 266, a stretch of 240 aa. The residues N-terminal to the proline-rich region appear to form a signal peptide, with a hydrophobic core, VLILLL. Consensus patterns for signal peptide cleavage sites (von Heijne, 1983) suggest that cleavage occurs between residues 24 and 25 predicting an N-terminal sequence for the mature protein of CPYCPYPSPK.

Fig. 33 Amino acid composition of the two regions of PP590 predicted protein (without signal peptide).

	proline-rich region		C-terminal region	
	residues	mole %	residues	mole %
alanine	2	0.8	7	8.0
cysteine	4	1.7	8	9.1
aspartate	0	0	5	5.7
glutamate	1	0.4	1	1.1
phenylalanine	5	2.1	1	1.1
glycine	0	0	9	10.2
histidine	14	5.9	1	1.1
isoleucine	8	3.3	8	9.1
lysine	17	7.1	5	5.7
leucine	4	1.7	15	17.0
methionine	0	0	0	0
asparagine	0	0	2	2.3
proline	112	46.9	6	6.8
glutamine	1	0.4	4	4.5
arginine	0	0	1	1.1
serine	6	2.5	4	4.5
threonine	12	5.0	5	5.7
valine	44	18.4	6	6.8
tryptophan	0	0	0	0
tyrosine	9	3.8	0	0

The predicted aa composition of the region between residues 26-266 of the PP590 predicted protein (fig. 33) confirms its proline rich nature, with valine, lysine, histidine and threonine making up 36% out of the remaining 53% of residues. The main body of the sequence is

composed of repeats of the three peptides given below (with their frequency of occurrence):

P	P	T/I/V/A	V/H	K/H/F	12x
P	P	H/Y	V/Y/I	P K/L	13x
P	P	V	V	P V T	5x

The repeats are located in the central region of the proline-rich sequence, accounting for 72% of the residues. However, there is no discernible order to their distribution within this region.

Hydrophobic region

The 86 residue C-terminal region of the PP590 encoded polypeptide has an overall hydrophobic composition (fig. 34) due to its high levels of leucine, isoleucine and alanine (fig. 33). When the NBRF protein database was searched using this region, a match to a soybean hydrophobic protein (Odani *et al.*,1987) was found.

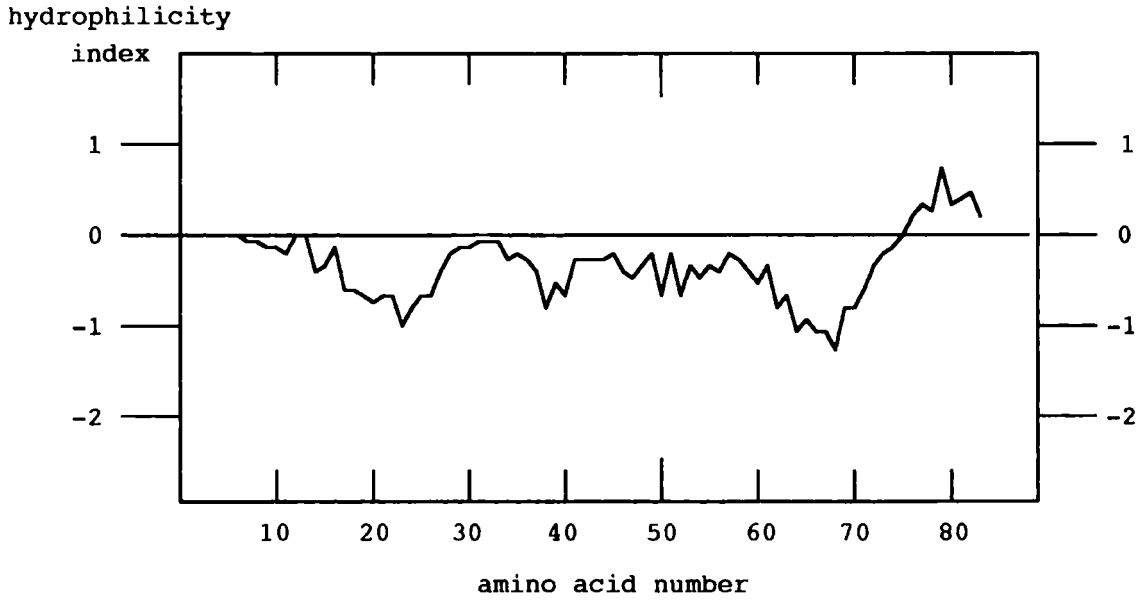


Fig. 34 Hopp Woods hydrophilicity plot for C-terminal region of PP590. Amino acids are numbered from residue 267 of the sequence

The optimised alignment (fig. 42) has a 34% identity and 69% homology when conservative aa substitutions are included (Pearson & Lipman, 1988).

PP590 flanking regions

The region including the start codon of PP590 conforms to the consensus for a functional translation start (Kozak, 1986). A consensus TATA box (Joshi, 1987) is located 80bp 5' of the start codon. 943bp were sequenced 5' to the start codon, although this region is predominantly A/T rich there are several short G/C rich sequences which are marked on fig. 32. At the 3' end of the ORF 1235bp have been sequenced. A sequence, GAATAA, occurs 154bp 3' from the stop codon, which conforms to the polyadenylation signal consensus in plants A/GAATA$(A)_{1-3}$ (Messing *et al.*,1983). Additional putative polyadenylation sites occur further 3', including the multiple overlapping type found in some plant genes (Lycett *et al.*,1983b).

Hybridisation of PP590 to genomic DNA

Genomic blots with DNA from PP and FF lines were probed with the 1kbp *Hind* III fragment containing almost all the coding sequence of PP590. Initial washing was to 2xSSC at 65°C, then subsequently to 0.1xSSC at 65°C. After high stringency washing (fig. 35a) a single band remained hybridised to by the PP590 probe in the PP DNA digested with *Hind* III and *Xba* I. The sizes of these fragments (1.0kbp & 4.7kbp respectively) correspond to those containing the gene in the genomic clone (fig. 31). In the *Eco* RI digest of PP DNA, a band at 3.5 kbp was hybridised to strongly and faint bands could be seen at 2.6kbp and 3.9kbp. The 3.9 and 3.5kbp bands represent partial digestion products and the 2.6kbp band the fragment predicted by complete digestion of the

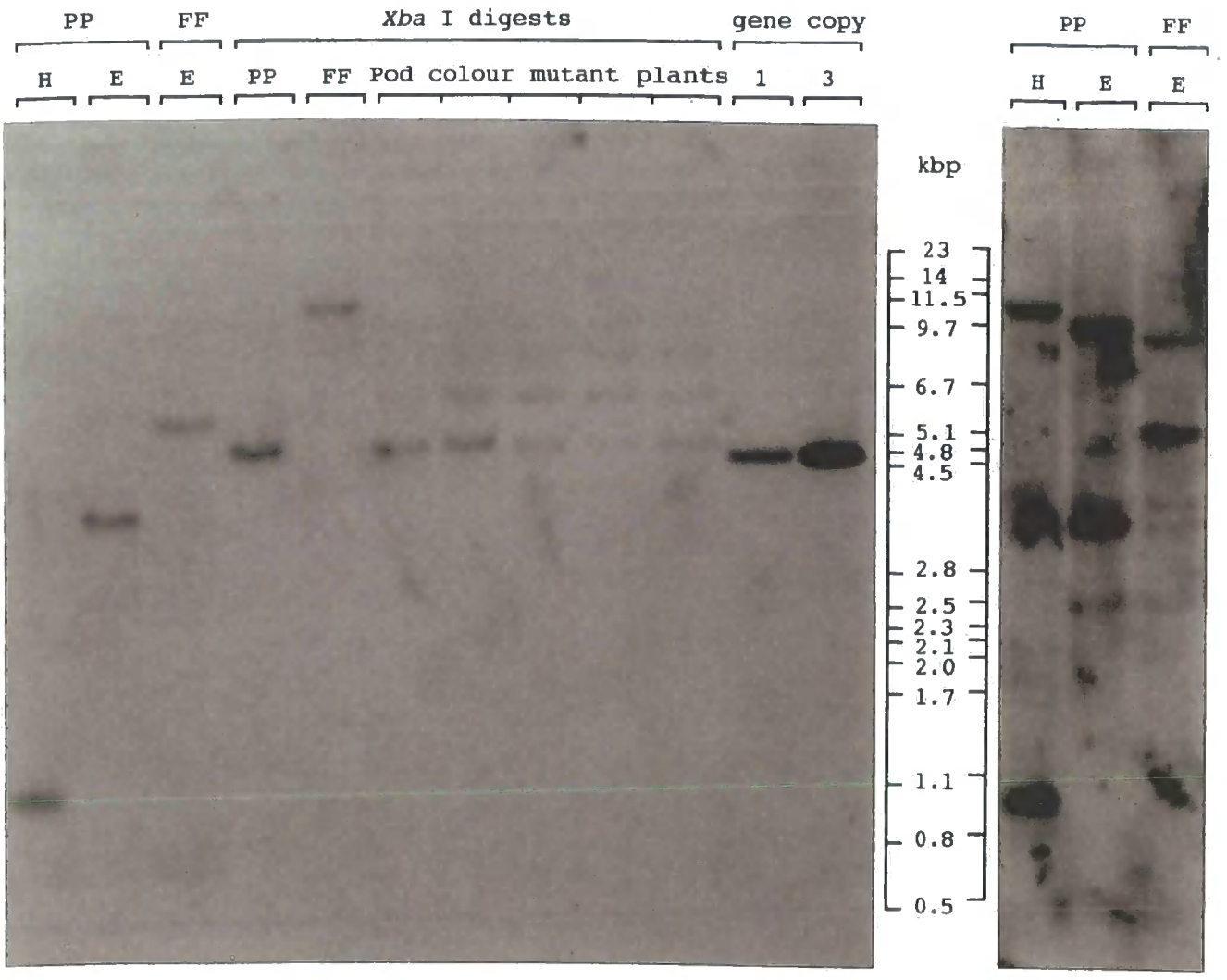


Fig. 35 Genomic digests of DNA from Purple-Podded (PP), mutant purple-podded and Feltham First (FF) plants probed with PP590. Rstriction enzymes used were; *Eco* RI (E), *Hind* III (H) and *Xba* I, gene copy equivalent amounts of PP590 were also run. a) shows the blot after washing to 0.1xSSC and b) after washing to 2xSSC. The position DNA size markers migrated to on the original gel is indicated.

genomic clone. The correspondence of the size of bands hybridised to on this blot and those predicted by the restriction map of the genomic clone demonstrates that no rearrangements have occurred during the cloning and subcloning procedures.

The intensity of the bands compared to that of gene equivalent amounts of PP590 indicate that in the PP line a single copy of the PP590 gene is present per haploid genome. A single copy of the PP590 gene is also present in the FF genome although the band hybridised to was of different size to that in the PP digests (5.1kbp *Eco* RI and 10kbp *Xba* I). Tracks containing *Xba* I digested genomic DNA from single PP plants exhibiting instability in the purple podded phenotype (green and purple or totally green pods) contained a range of faint bands (4.9, 6.0, 7.8 & 9.4kbp), the sizes of which could be accounted for by partial restriction fragments predicted by the genomic clone.

When this genomic blot had been washed to lower stringency (fig. 35b) one additional band was hybridised to at single gene copy level in all digests of all DNAs. This extra fragment was of different sizes in the PP and FF digests (*Eco* RI fragments of 9.5 & 9.0kbp respectively). There were no partial digestion products associated with this extra band in any digests and the DNA appeared to have digested to completion when the gel was stained and photographed before blotting (data not shown).

pPPL927 positive genomic clone

A single transformant, designated PP927, was purified from the genomic library screen with pPPL927 insert. Using the same strategy employed for PP590, a restriction map for the genomic clone was deduced and this is shown in fig. 36. The genomic fragment cloned is 15kbp in length and the region homologous to pPPL927 lies adjacent to the left arm of the λ vector, within a 2.7kbp fragment, extending from the junction with the vector to a *Hind* III site within the genomic DNA. This fragment was subcloned into plasmid vector, restriction mapped and regions from this subcloned into M13 and subjected to automated DNA sequencing. A detailed restriction map and sequencing strategy are also shown on fig. 36. 2791bp. were sequenced (in both directions and all restriction sites sequenced through) from the *Sau* 3A site used to cut the genomic DNA to the *Hind* III site within the genomic clone, this sequence is presented in fig. 37.

A pectinesterase-like gene: Coding sequence

An ORF encoding a polypeptide with homology to tomato pectinesterase (Ray *et al.*,1988) was deduced, its start codon is 440bp from the junction of the genomic clone and the λ arm, and the stop codon is 2185bp from the same point. The coding sequence is interupted by an intron of 84bp which occurs at the same point in the aa sequence as that in the pectinesterase-like gene isolated from *Brassica napus* (Albani *et al.*,1991). This results in two exons of 967bp and 695bp with a total coding sequence length of 1662bp, encoding 554 aa.

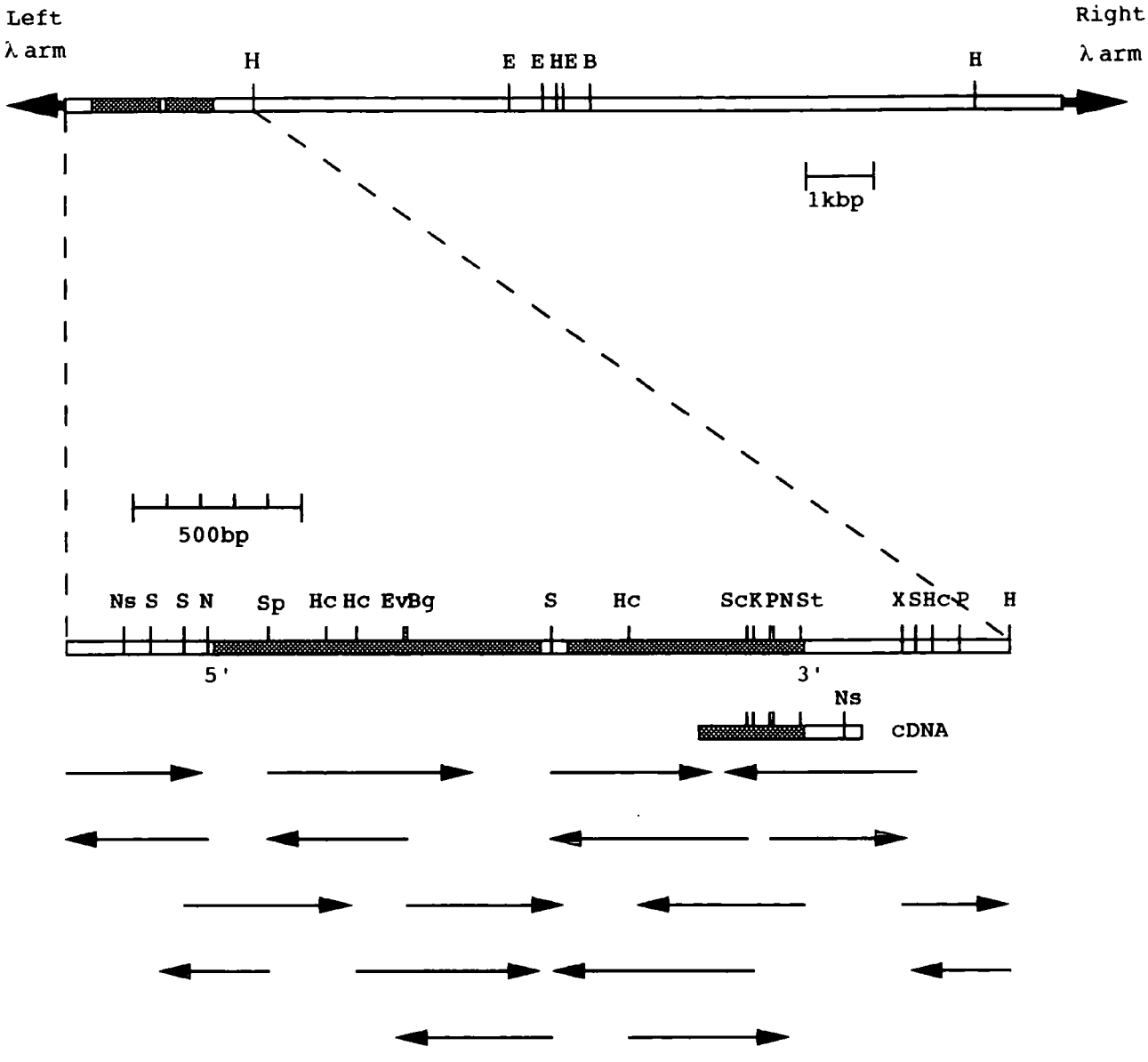
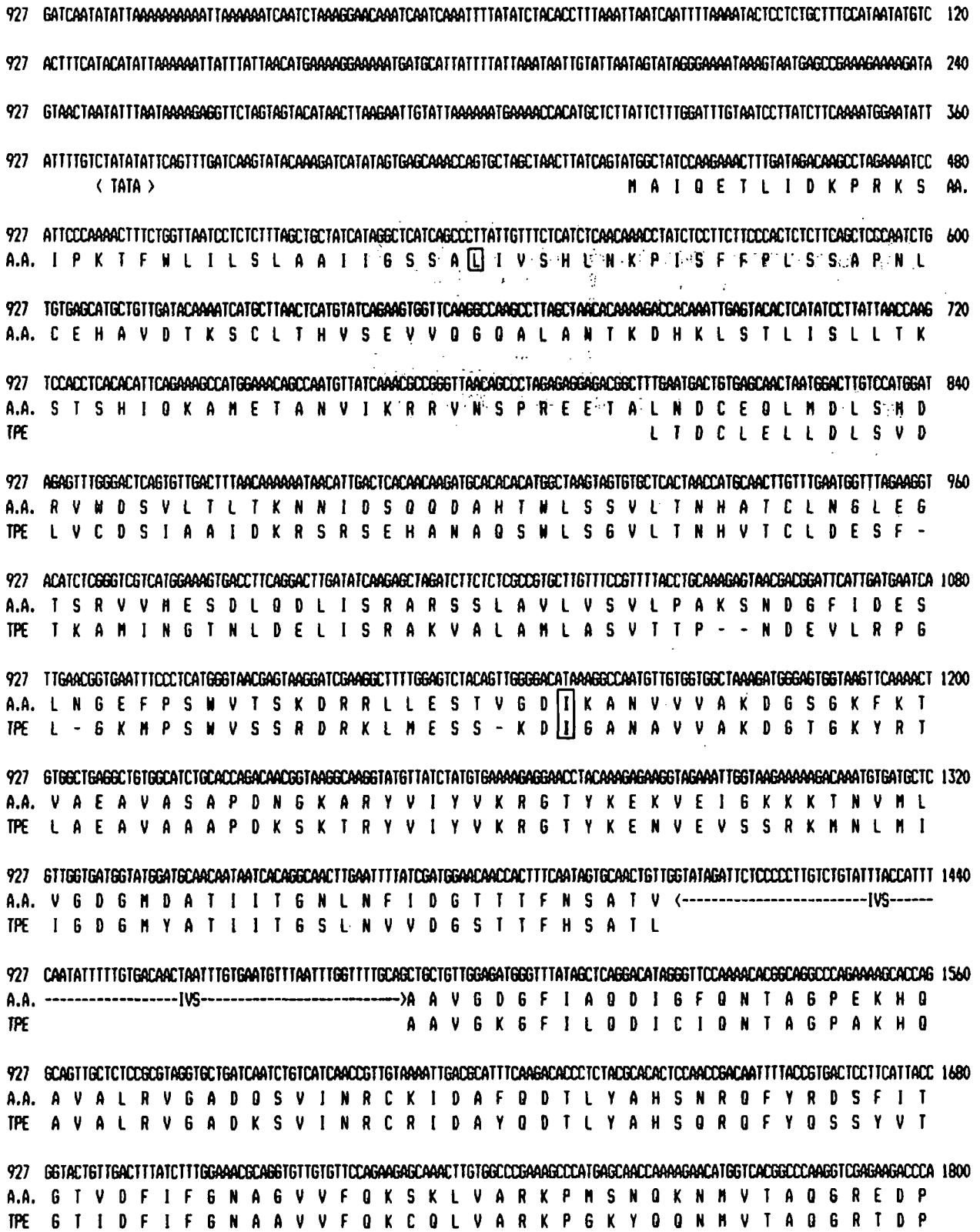


Fig. 36 Restriction maps of PP927 genomic clone, the region sequenced from it and pPPL927 cDNA. Restriction enzymes are abbreviated as: B, *Bam* HI; Bg, *Bgl* II; E, *Eco* RI; Ev, *Eco* RV; H, *Hind* III; Hc, *Hinc* II; K, *Kpn* I; N, *Nhe* I; Ns, *Nsi* I; P, *Pst* I; S, *Sst* I; Sc, *Sca* I; Sp, *Sph* I; St, *Stu* I; X, *Xba* I. The coding region is highlighted and arrows below the sequenced region represent individual sequencing runs.


927 GATCAATATATTAAAAAAAAAATTAAAAAATCAATCTAAAGGAACAAATCAATCAAATTTTATATCTACACCTTTAAATTAATCAATTTTAAAATACTCCTCTGCTTTCCATAATATGTC 120
927 ACTTTCATACATATTAAAAAATTATTTATTAACATGAAAAGGAAAAATGATGCATTATTTTATTAAATAATTGTATTAATAGTATAGGGAAAATAAAGTAATGAGCCGAAAGAAAAGATA 240
927 GTAACTAATATTTAATAAAAGAGGTTCTAGTAGTACATAACTTAAGAATTGTATTAAAAAATGAAAACCACATGCTCTTATTCTTTGGATTTGTAATCCTTATCTTCAAAATGGAATATT 360
927 ATTTTGTCTATATATTCAGTTTGATCAAGTATACAAAGATCATATAGTGAGCAAACCAGTGCTAGCTAACTTATCAGTATGGCTATCCAAGAAACTTTGATAGACAAGCCTAGAAAATCC 480
〈 TATA 〉 M A I Q E T L I D K P R K S AA.
927 ATTCCCAAAACTTTCTGGTTAATCCTCTCTTTAGCTGCTATCATAGGCTCATCAGCCCTTATTGTTTCTCATCTCAACAAACCTATCTCCTTCTTCCCACTCTCTTCAGCTCCCAATCTG 600
A.A. I P K T F W L I L S L A A I I G S S A L I V S H L N K P I S F F P L S S A P N L
927 TGTGAGCATGCTGTTGATACAAAATCATGCTTAACTCATGTATCAGAAGTGGTTCAAGGCCAAGCCTTAGCTAACACAAAAGACCACAAATTGAGTACACTCATATCCTTATTAACCAAG 720
A.A. C E H A V D T K S C L T H V S E V V Q G Q A L A N T K D H K L S T L I S L L T K
927 TCCACCTCACACATTCAGAAAGCCATGGAAACAGCCAATGTTATCAAACGCCGGGTTAACAGCCCTAGAGAGGAGACGGCTTTGAATGACTGTGAGCAACTAATGGACTTGTCCATGGAT 840
A.A. S T S H I Q K A M E T A N V I K R R V N S P R E E T A L N D C E Q L M D L S M D
TPE L T D C L E L L D L S V D
927 AGAGTTTGGGACTCAGTGTTGACTTTAACAAAAAATAACATTGACTCACAACAAGATGCACACACATGGCTAAGTAGTGTGCTCACTAACCATGCAACTTGTTTGAATGGTTTAGAAGGT 960
A.A. R V W D S V L T L T K N N I D S Q Q D A H T W L S S V L T N H A T C L N G L E G
TPE L V C D S I A A I D K R S R S E H A N A Q S W L S G V L T N H V T C L D E S F -
927 ACATCTCGGGTCGTCATGGAAAGTGACCTTCAGGACTTGATATCAAGAGCTAGATCTTCTCTCGCCGTGCTTGTTTCCGTTTTACCTGCAAAGAGTAACGACGGATTCATTGATGAATCA 1080
A.A. T S R V V H E S D L Q D L I S R A R S S L A V L V S V L P A K S N D G F I D E S
TPE T K A M I N G T N L D E L I S R A K V A L A M L A S V T T P - - N D E V L R P G
927 TTGAACGGTGAATTTCCCTCATGGGTAACGAGTAAGGATCGAAGGCTTTTGGAGTCTACAGTTGGGGACATAAAGGCCAATGTTGTGGTGGCTAAAGATGGGAGTGGTAAGTTCAAAACT 1200
A.A. L N G E F P S W V T S K D R R L L E S T V G D I K A N V V V A K D G S G K F K T
TPE L - G K M P S W V S S R D R K L M E S S - K D I G A N A V V A K D G T G K Y R T
927 GTGGCTGAGGCTGTGGCATCTGCACCAGACAACGGTAAGGCAAGGTATGTTATCTATGTGAAAAGAGGAACCTACAAAGAGAAGGTAGAAATTGGTAAGAAAAAGACAAATGTGATGCTC 1320
A.A. V A E A V A S A P D N G K A R Y V I Y V K R G T Y K E K V E I G K K K T N V M L
TPE L A E A V A A A P D K S K T R Y V I Y V K R G T Y K E N V E V S S R K M N L M I
927 GTTGGTGATGGTATGGATGCAACAATAATCACAGGCAACTTGAATTTTATCGATGGAACAACCACTTTCAATAGTGCAACTGTTGGTATAGATTCTCCCCCTTGTCTGTATTTACCATTT 1440
A.A. V G D G M D A T I I T G N L N F I D G T T T F N S A T V 〈----------IVS------
TPE I G D G M Y A T I I T G S L N V V D G S T T F H S A T L
927 CAATATTTTTGTGACAACTAATTTGTGAATGTTTAATTTGGTTTTGCAGCTGCTGTTGGAGATGGGTTTATAGCTCAGGACATAGGGTTCCAAAACACGGCAGGCCCAGAAAAGCACCAG 1560
A.A. ----------IVS----------〉A A V G D G F I A Q D I G F Q N T A G P E K H Q
TPE A A V G K G F I L Q D I C I Q N T A G P A K H Q
927 GCAGTTGCTCTCCGCGTAGGTGCTGATCAATCTGTCATCAACCGTTGTAAAATTGACGCATTTCAAGACACCCTCTACGCACACTCCAACCGACAATTTTACCGTGACTCCTTCATTACC 1680
A.A. A V A L R V G A D Q S V I N R C K I D A F Q D T L Y A H S N R Q F Y R D S F I T
TPE A V A L R V G A D K S V I N R C R I D A Y Q D T L Y A H S Q R Q F Y Q S S Y V T
927 GGTACTGTTGACTTTATCTTTGGAAACGCAGGTGTTGTGTTCCAGAAGAGCAAACTTGTGGCCCGAAAGCCCATGAGCAACCAAAAGAACATGGTCACGGCCCAAGGTCGAGAAGACCCA 1800
A.A. G T V D F I F G N A G V V F Q K S K L V A R K P M S N Q K N M V T A Q G R E D P
TPE G T I D F I F G N A A V V F Q K C Q L V A R K P G K Y Q Q N M V T A Q G R T D P

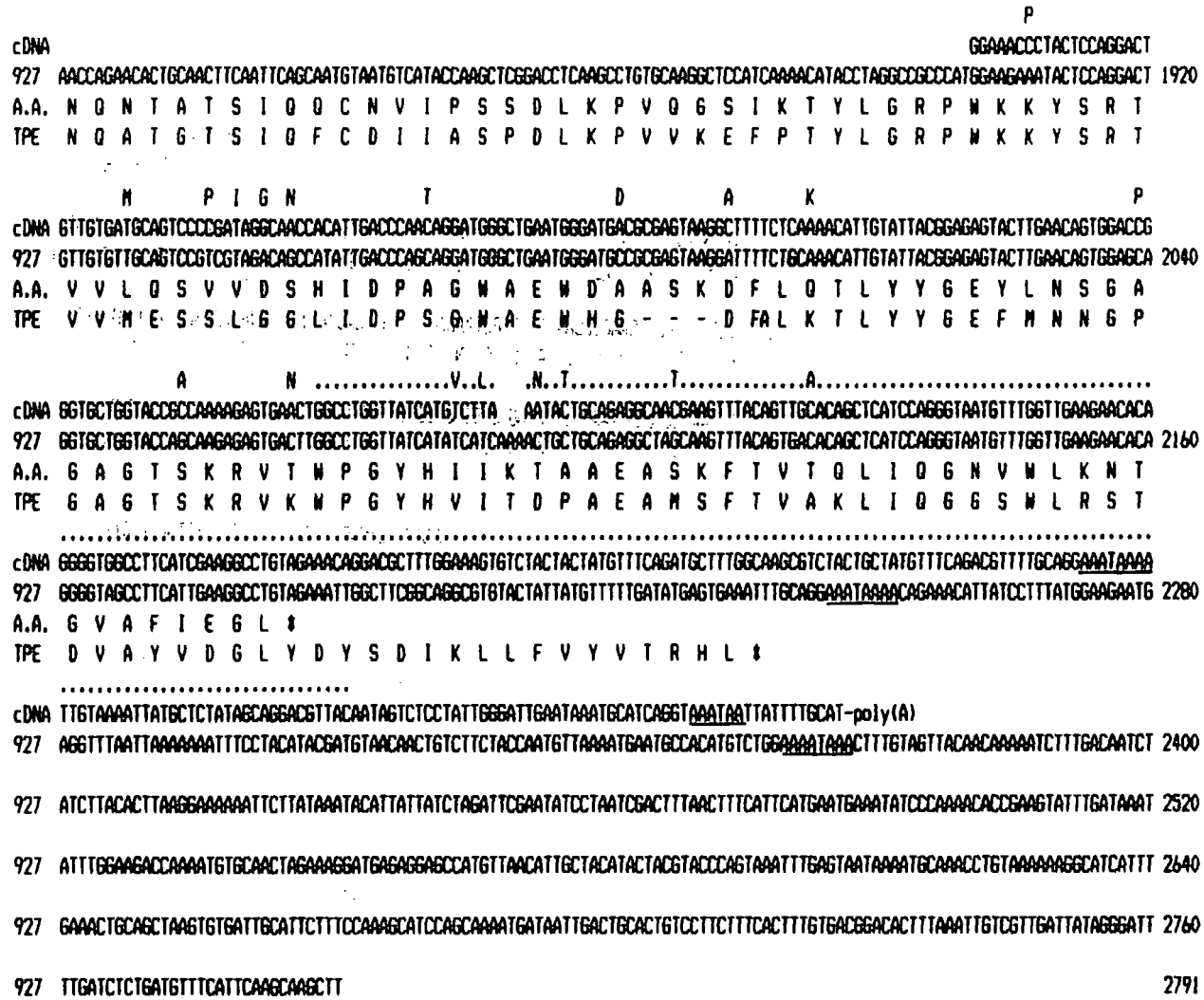


Fig. 37 Nucleotide (927) and derived amino acid sequence (A.A.) of PP927 pectinesterase-like gene. Nucleotide sequence of the pod expressed cDNA, pPPL927, (cDNA) is shown above the genomic sequence, with deviations from the protein sequence derived from genomic DNA sequence shown above. The extent of the plasmid cDNA, pPP927 is indicated thus ..... and consensus polyadenylation signals are underlined. Derived amino acid sequence from the tomato pectin esterase cDNA (Ray et al.,1988) is also presented (TPE) with gaps introduced for homology. The N-terminal residue of the mature pectinesterase protein (Markovic and Jornvall, 1986) and N-terminal residue predicted for PP927 are boxed.

The flanking sequences

A match to the central region of the consensus TATA box sequence in plants, TCACTATATATAG (Joshi, 1987) is located 65bp from the start of the ORF, CTATATAT. The start codon of the ORF lies within a sequence compatible with translation initiation (Kozak, 1986). No other regions of interest could be found in the 5' flanking region and comparison of this area with the 5' flanking region from the other plant pectinesterase-like gene sequenced (from *Brassica napus*) yielded no sequences with significant homology. There are two consensus polyadenylation signals 60 and 170bp downstream from the stop codon.

Comparison with the cDNAs

The genomic sequence shows significant differences when compared with that of the cDNA used to isolate it, there are 36 base changes in the 281bp coding region overlap, resulting in 19 aa substitutions. There is also a deletion of a single codon from the cDNA with respect to the gene. The two 3' flanking regions diverge completely after the stop codon with the only homologous region being that including the first polyadenylation signal in both sequences, GCAGGAAATAAAA.

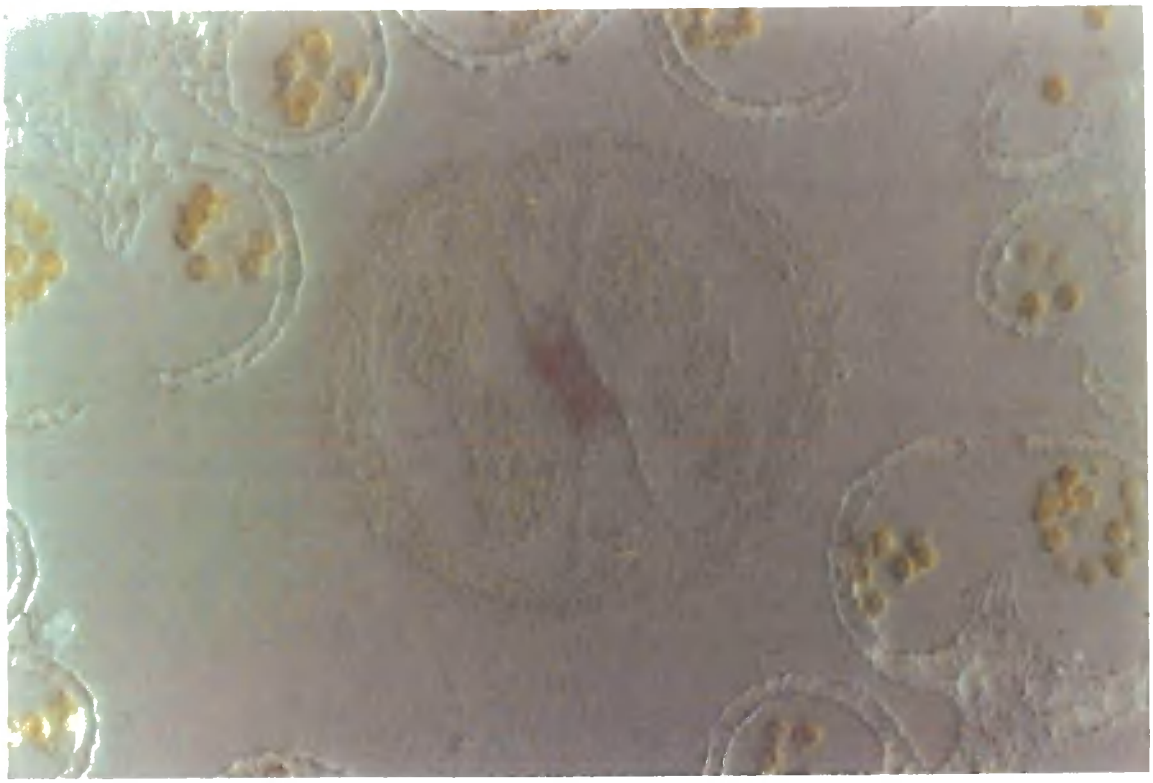


Fig. 38 *In situ* hybridisation of PP927 to *Arabidopsis* flower. The photograph shows a transverse section (at 20x magnification) through an Arabidopsis flower prior to pollen dehiscence. Hybridisation by the PP927 probe results in a red colouration, this is concentrated in the newly fused septum in the gynoecium (centre). Pollen sacs (and pollen) can be seen surrounding the gynoecium

***In situ* hybridisation studies with PP927**

PP927 was digoxygenin labelled and used as a probe on sections through developing *Arabidopsis thaliana* flowers. Hybridised probe was detected with alkaline phosphatase labelled anti-digoxygenin antibodies and visualised using napthol AS-MX phosphate and fast red TR dye. The probe hybridised to mRNA in cells of the tapetum of developing anthers prior to pollen dehiscense and strongly to the central area of the septum after the fusion of its two halves (fig. 38). PP927 homologous mRNA remained present in the central area of the septum after fertilization. Low levels of hybridisation were also seen in petal cells and identical results were obtained when tomato pectinesterase DNA was used as a probe (Spence,J., personal communication).

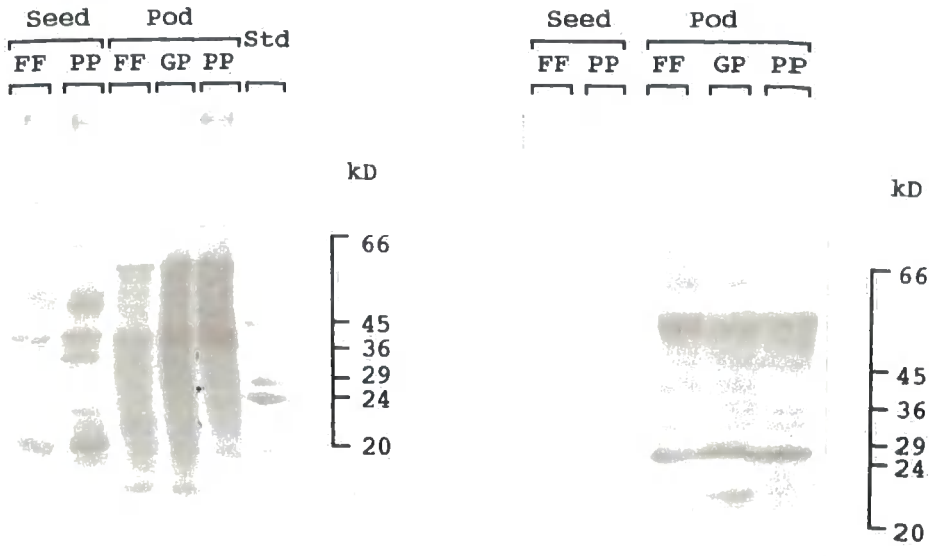


Fig. 39 Polyacrylamide gel (left) of total protein extracted from seed and pods from PP, GP and FF plants. W Blot (right) of the duplicate half of the same gel probed with anti-(tomato fruit) pectinesterase antibodies. The Mrs of the standard proteins on the gel are indicated and the positions these would have occurred at on the filter are shown.

Immunological studies with anti-pectinesterase antibodies

Antiserum raised against tomato pectinesterase was used to probe western blots of polyacrylamide protein gels run under reducing conditions. These gels contained proteins extracted from 5 daf pods from FF, PP and GP, extracts of protein from mature FF and PP cotyledons were also run. The results (fig. 39 and others not presented) are somewhat ambiguous due to the high background and unknown specificity of the antiserum used. A protein of 28k Mr in all three pod extracts reacts strongly with the antiserum and a protein of 34k Mr reacts in the seed extracts. Running the gels under non-reducing conditions had no effect on the sizes of these bands. Probing a blot with phosphatase substrate and no antiserum or secondary antibody failed to account for any of the background being due to phosphatases in the extracts.

Hybridisation of PP927 to genomic DNA

Genomic Southern blots with digests of PP and FF genomic DNA were prepared and probed with the 1.8kbp *Eco* RV to *Hind* III fragment of PP927. This probe includes all of the region encoding the pectinesterase-like sequence and 146bp 5' to this, ie. extending into the sequence encoding the N-terminal region. Washing was initially to 1xSSC at 65°C then to 0.1xSSC at 65°C. Results from this and the blot probed with the 5' end of PP927 are shown in fig. 40 and tabulated in fig. 41.

The bands remaining hybridised to by this probe at high stringency (fig. 40a) were: In the *Nsi* I digested PP DNA 3.0 and 4.5kbp; in the *Eco* RV digest of both pea lines, 13, 7.0 and 6.2kbp, plus an extra band at 6.7kbp in the FF DNA; and in the *Eco* RI digests 7.2 and 4.5kbp. Both the *Nsi* and *Eco* RV digests of PP contain DNA fragments, hybridised to by this probe, of identical length (4.5 and 6.2kbp respectively) to those hybridised to in similar digests of the genomic clone (data not shown). That these internal fragments of the genomic clone are identical to those found in genomic DNA shows that no rearrangements have occurred during the cloning and screening procedures.

The intensity and number of the bands hybridised to in each digest at high stringency, when compared to the gene copy equivalent digests, indicate that there are three highly homologous genes in the genomes of both the PP and FF lines and a further highly homologous gene in the FF line. At least two extra bands could be seen in each digest of both lines after washing at lower stringency (fig. 40a), suggesting that other genes with similar sequence are present in the genome of both lines.

A similar genomic blot was probed with the region 5' to the *Eco* RV site, encoding only the non-pectinesterase-like (N-terminal) region of

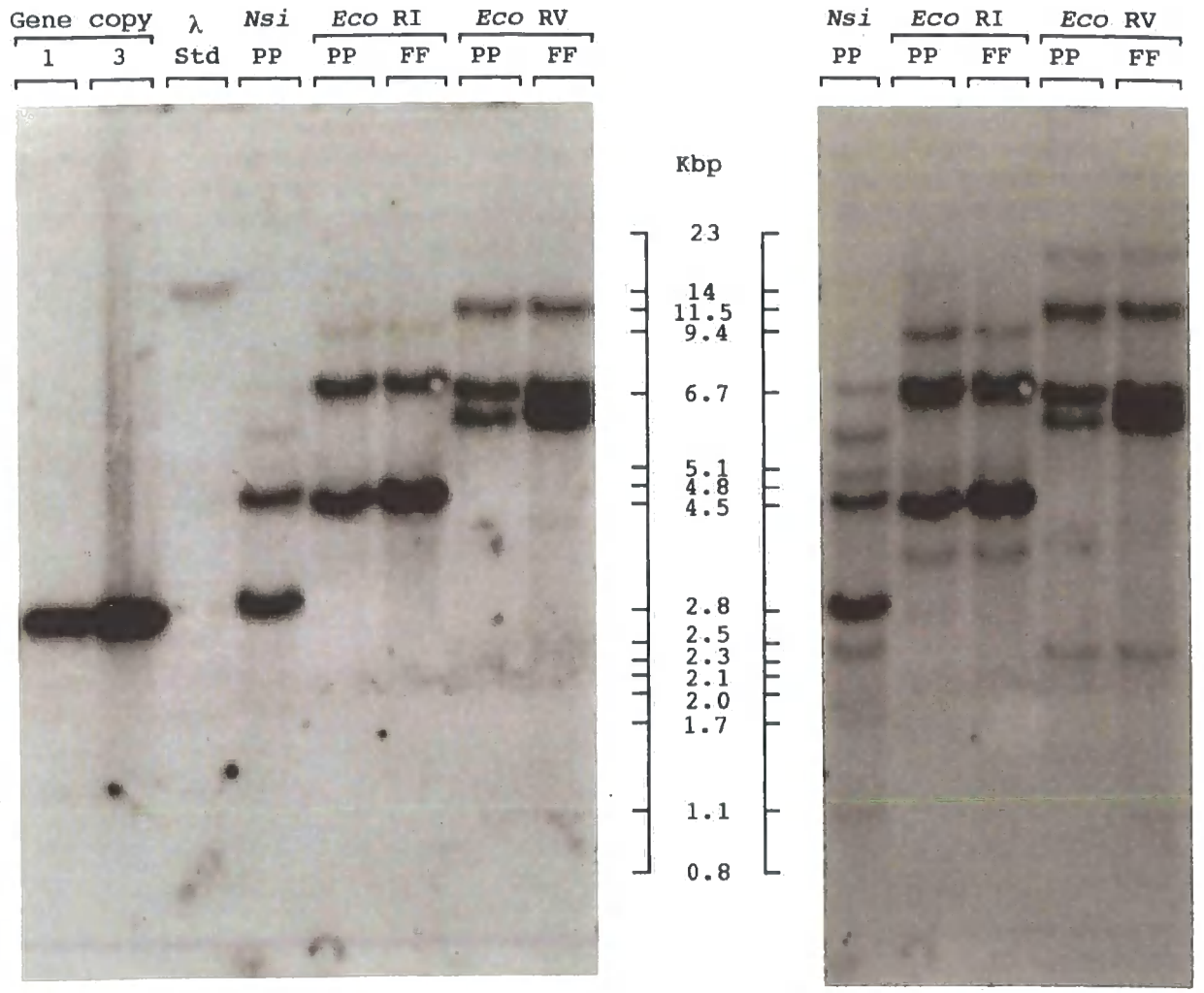


Fig. 40a Genomic digests of FF and PP DNA restricted with *Nsi* I, *Eco* RI or *Eco* RV probed with the 3' (pectinesterase encoding) region of PP927. The filter was first washed to 2xSSC and exposed (right) and subsequently to 0.1xSSC (left). Gene copy equivalents of PP927 were also run and the position standard DNA bands migrated to on the original gel is marked.

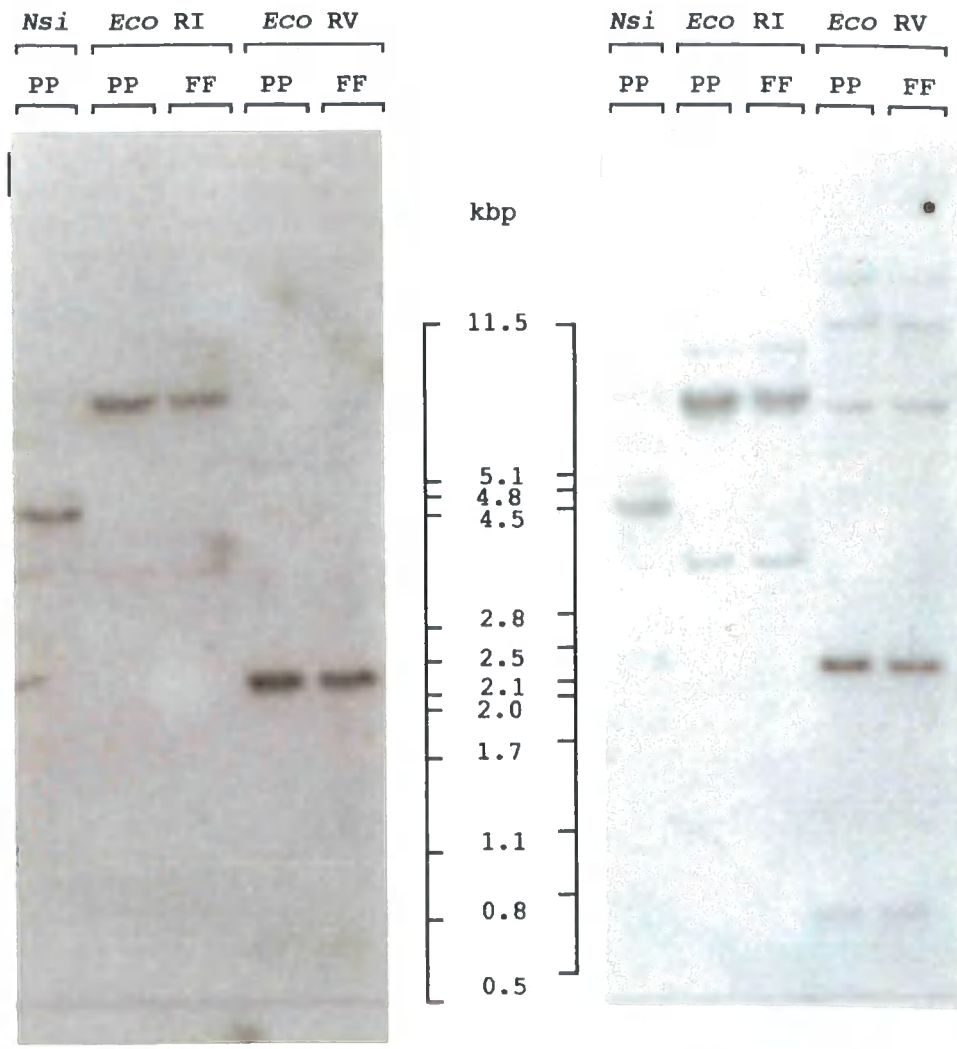


Fig. 40b Genomic digests of Purple-Podded (PP) and Feltham First (FF) DNA digested with *Eco* RI, *Eco* RV and *Nsi* I probed with the region of PP927 5' of the *Eco* RV site. The filter was washed to 1xSSC (right) and subseuently to 0.1xSSC (left). The position DNA size marker bands ran to on the original gel is indicated.

Digest:	*Nsi* I				*Eco* RI				*Eco* RV			
Probe:	5'		3'		5'		3'		5'		3'	
Wash:	High	Low	High	Low	High	Low	High	Low	High	Low	High	Low
		7.3		7.2		9.5		9.5		19		19
		5.8		5.8	7.2	7.2	7.2	7.2		13	13	13
		5.0		5.0		7.0		7.0		7.0	7.0	7.0
	4.5	4.5	4.5	4.5			4.5	4.5			6.2	6.2
			3.0	3.0		3.5		3.5	2.3	2.3		2.3
		2.4		2.4						0.7		
				1.1								

Fig. 41 Sizes of genomic DNA fragments (kbp) from various digests of PP DNA, hybridised to by DNA probes encoding the two regions of the protein predicted by PP927. The 5' probe encodes the N-terminal region and the 3' probe the C-terminal pectinesterase-like region. Stringency of washing is denoted by High - 0.1xSSC and Low - 1xSSC. Fragment sizes underlined are those corresponding to internal fragments of the genomic clone or, in the case of *Eco* RI and the 5' *Eco* RV fragment, those deduced to contain the PP927 gene (see text).

the predicted protein. When washed to high stringency (0.1xSSC at 65°C), only one band remained hybridised to in all digests, of the same size in both FF and PP lines. The size of the band in the *Nsi* digest (4.5kbp) corresponds to the internal fragment from the genomic clone (see above). The band hybridised to in the *Eco* RI digests corresponds to one hybridised to by the *Eco* RV-*Hind* III probe (7.2kbp), this fragment must be the one which contains the PP927 gene (as the other at 4.5kbp is too short to contain the 6.5kbp *Sau* 3A-*Eco* RI fragment carrying the gene in the genomic clone Fig. 36). In the *Eco* RV digest a 2.3kbp band was hybridised to by this (5') probe which presumably encompasses this *Sau* 3A to *Eco* RV probe.

When washed at lower stringency (1xSSC at 65°C) at least three additional bands in each digest were hybridised to in the genomic DNA

(see fig. 40b) with no differences apparent between the FF and PP lines. All the fragments hybridised to by this probe (except for a 0.7kbp *Eco* RV fragment) were also hybridised to by the 3' probe.

CHAPTER THIRTEEN: PURPLE PODDED PHENOTYPE - DISCUSSION

cDNA Screens

The initial strategy used to identify sequences specific to purple pod tissue was to carry out differential screening of a cDNA library. To this end a plasmid cDNA library was constructed, although unfortunately the insert size was generally small (200-500bp), despite numerous attempts to create a library containing larger inserts using differing protocols. The poly(A) enriched RNA used (again prepared several times) did not appear to be degraded, as judged by lack of smearing of bands hybridised to by probes on blots of gels (data not shown). A seed cDNA library was constructed which contained clones with inserts 400-2000bp in size using the same methods. Possibly the pod RNA isolated contains a contaminant which reduces the efficiency of cDNA synthesis.

To test whether cDNAs encoding enzymes likely to be affecting the purple colour production could be isolated from the plasmid library, a trial screen was conducted. DNA encoding chalcone synthase (CHS) and the maize *A1* gene encoding dihyroflavonol-4-reductase (Reddy *et al.*,1987) were used. CHS cDNAs were isolated (see below) but no *A1* homologous cDNAs. These probes were not continued with as when used on northern blots containing RNAs from 5 daf PP, GP and FF pods, mRNA levels were the same in all lines. A similar result was also obtained using chalcone isomerase as a probe (data not shown). These results suggest that the lack of pigment in the pods from GP and FF lines is not the non-functioning of one of these enzymes. It is possible that a mutation causing a frame shift or codon change to a stop could occur without altering the size of an mRNA. However, studies in this laboratory on a pea legumin gene with such a mutation, expression studies on a mutated soybean trypsin inhibitor gene and plant transformation with a soybean lectin pseudogene with in-frame stop codons, demonstrate that such mRNAs

fail to accumulate (Thompson *et al.*,1991, Jofuku *et al.*,1989, Voelker *et al.*,1990). The identical levels of chalcone synthase mRNA in both green and purple tissue also shows that the mutation does not affect the *a* or *a2* loci, as these are required to be functional for maximal levels of CHS mRNA to accumulate in petals (Harker *et al.*,1990).

As the CHS screen had proved successful, a differential screen was performed using FF cDNA as the "green" line (insufficient GP material was available to make poly(A) enriched RNA from). Three differentially expressed cDNAs were isolated and sequenced (see below), all containing short (<400bp) inserts. In an attempt to obtain longer cDNAs and therefore increase the efficiency of hybridisation of the probes to a genomic library, a λ cDNA library was constructed. A slightly longer (478bp) clone identical to pPP927 was obtained but still not full-length and no clones were isolated with the other probes. In retrospect, a better but more time consuming strategy would have been to screen the λ cDNA library with the probes separately, as the mixed probe certainly seems to have been biased in favour of pPP927. However, at this stage it was decided to proceed to the genomic screen with the cDNAs isolated.

Chalcone synthase cDNAs

Two differing types of CHS cDNAs were isolated, one (pPP166 and 888) shows close homology (67% identity) to the CHS sequence encoded by *Phaseolus vulgaris* cDNAs (Ryder *et al.*,1987), while the other (pPP372) is less homologous (56% identity). Recently, the sequences of three full length pea CHS cDNAs have been published (Ichinose *et al.*,1992) which include the regions covered by the cDNAs reported here. These cDNAs were isolated from a library prepared from epicotyl tissue after treatment with fungal elicitor. Although the encoded aa sequences from these are homologous to the pod cDNAs none of them is identical, the closest, pCC2

has 72% identity to pPP166 which is slightly less than the identity between the two pod cDNA species (76%). Due to the shortness of the pod cDNAs and the small difference in homologies it is not possible to enlarge upon the observation these authors (Ichinose *et al.*,1992) make, that CHS clones from elicitor treated libraries are structurally more similar to each other than to the one clone previously isolated from an untreated library (Harker *et al.*,1990).

CHS is an important regulatory enzyme, it catalyzes the first step in the flavonoid specific biosynthesis pathway (fig. 2, Ebel & Hahlbrock, 1982, Heller & Forkmann, 1988). Mutations of the genes encoding this enzyme can cause blockage of this pathway and resultant lack of pigmentation (Coen & Carpenter, 1988, Wienand *et al.*,1986). Northern blots suggest that this is not the case in the green/purple pod mutation (see above) and extracts from GP and FF pods were found to contain flavonoids when run on thin layer chromatograms (data not shown).

In *Phaseolus, Petunia*, soybean and pea, CHS is a member of a differentially expressed, multigene family (Ryder *et al.*,1987, Koes *et al.*,1989, Wingender *et al.*,1989, Harker *et al.*,1990). Results from probing genomic DNA (digested with *Eco* RI) from the PP, GP and FF lines with *Phaseolus* CHS cDNA agree with those published (Harker *et al.*,1990). Eight fragments are hybridised to by this probe at low stringency with no size differences between the PP and GP lines, only four of the fragments were, however, conserved between FF and the other lines (data not shown).

pPP812 cDNA

The pPP812 cDNA hybridises to a small (500b) mRNA in the pods from the PP line. No hybridisation could be detected to the FF RNA and

evidence from genomic blots (not shown) suggests that this sequence is absent from this line. This result together with the decreased level of mRNA in the GP line make the lack of success in isolating a longer cDNA or genomic clone disappointing. Evidence from northern blots containing RNA from coloured petals suggest that the gene product from pPP812 is not involved in flavonoid biosynthesis as no hybridisation could be detected (fig. 30).

Genomic library

A PP genomic library was constructed containing $9x10^5$ distinct transformants. The probability of obtaining a given DNA sequence from the pea genome in a library of this size is 0.94, assuming an average genomic clone length of 15kbp (Clarke & Carbon, 1976). On this basis it was decided to proceed and screen the library.

Clones homologous to two of the cDNA probes were isolated from the genomic library (see below), the third, pPP812, was unsuccessful, despite repetition of the screen. This cDNA had the smallest insert and lack of specificity of the probe may be the reason why no clones could be purified from apparently duplicating plaques in the initial screens.

PP590 Genomic clone proline-rich sequence

The nature of this sequence led to many similarly proline-rich but not necessarily homologous proteins being found when the NBRF protein database was searched. However, members of the family of repetitive hydroxyproline-rich proteins isolated from soybean cell walls (Hong *et al.*,1987, Averyhart-Fullard *et al.*,1988, Datta *et al.*,1989) appeared to be the best match. These have repeating units of P-P-V/I-Y-K or P-P-V-Y-K-P-P-V-E-K.

Repetitive proline-rich proteins

Until recently, repetitive proline-rich plant proteins had been divided into three classes (Showalter & Varner, 1989):

1) Arabinogalactans, rich in hydroxyproline, alanine, and serine, those peptides sequenced so far contain alternating alanine and hydroxyproline residues. These proteins are extensively glycosylated (to >90% of Mr), with branched side chains composed of galactose and arabinose.

2) Solanaceous lectins, composed of two domains, one proline and serine rich, glycosylated and possibly similar to extensins. The other domain is glycine and cysteine rich and crosslinked by disulphide bridges.

3) Extensins - see below.

Extensins

The most widely studied family of cell wall proteins are the extensins (Cassab & Varner, 1988, Varner & Lin, 1989), these are hydroyproline-rich glycoproteins characterised by repeating units composed of two parts; firstly S-P-P-P-P (which may occur on its own), followed by a unit of one to seven residues, composed mainly of valine, histidine, tyrosine, lysine, threonine, and proline. When the extensin peptides are isolated it is found that most of the prolines of the S-$(P)_4$ region are hydroxylated, and glycosylated, with arabinose tri- and tetramers predominating (Lamport, 1969). The serine residues are also glycosylated, with single galactose units (Lamport *et al.*,1973).

It is thought that the extensin molecules form flexible rods, with the S-$(P)_4$ units forming rigid domains of polyproline II helical structure stabilised by their carbohydrate component. The remaining residues of the repeat function as flexible spacers reponsible for intermolecular crosslinkage, possibly through the formation of isodityrosyl links (Epstein & Lamport, 1984, Kieliszewski & Lamport,

1986, Showalter & Varner, 1989). A model where extensin forms an independent, structural, crosslinked network, associated with the cellulose micro-fibrils has been proposed (Lamport & Epstein, 1983, Stafstrom & Staehelin, 1988, Kieliszewski & Lamport, 1988). Recently though, doubt has been expressed as to whether such a rigid network exists and could adapt to the needs of cell growth (Varner & Lin, 1989, Talbot & Ray, 1992).

In this laboratory two extensin genes have been isolated from *Brassica napus* (Evans *et al.*,1990, Gatehouse *et al.*,1990). When cDNA homologous to one of these genes was used to isolate extensin-like sequences from an *Arabidopsis thaliana* genomic library (Yaish, 1990), a gene was isolated with a single polyproline pentamer and a similar sequence to that of the C-terminal region of PP590 (see below).

A fourth class of repetitive proline-rich proteins

It is now emerging that a fourth class of repetitive proline-rich proteins exists (Marcus *et al.*,1991), and it is this class to which the proline-rich sequence of PP590 belongs. Proteins in this class are characterised by proline-rich repeating units but lack the $S-(P)_4$ units of the extensins. In addition to the repetitive proline-rich (RPRP) family from soybean, DNA sequences encoding proteins of similar composition have also been characterised in tissues from other dicot and monocot species: Carrot root, with repeating units of P-P-I/V-H-K and P-P-V-Y-T (Chen & Varner, 1985b); soybean root nodules, repeats of P-P-H/Y/L-E-K-P-P- $(X)_{3-4}$, where X=P,E,Y,Q or H (Franssen *et al.*,1987); and maize coleoptiles, repeats of P-P-T-Y-T-P-S-P-K-P-P-T-P-K-P-T or P-P-T-Y-T-P-S-P-K-P-P-A-T-K-P-P-T-P-K-P-T (Stiefel *et al.*,1988), with similar genes also found in *Sorghum* (Raz *et al.*,1991), rice (Caelles *et al.*,1992) and the gymnosperm, Douglas fir (Kielisewski *et al.*,1992).

Recently, the sequence of a gene encoding a proline-rich protein expressed in tomato fruit has been reported (Salts *et al.*,1991 & 1992), this contains a C-terminal region strikingly similar to that of PP590 (see below) and an N-terminal region of similar composition with repeating units of P-P-H/Y/I/V-V-K/H/S/Y and P-P-S/T/F/V-T-P-K.

Expression and role of class IV RPRPs.

The soybean RPRP family has been the most extensively studied, consisting of at least three members, differentially expressed in a limited number of cells in particular organs: 1) expressed strongly in the mature hypocotyl, root and immature seed coat; 2) highly expressed in the apical hypocotyl, germinating cotyledons and cultured cells; 3) most abundant in leaves, stem, pods and seed coat. mRNA from all three members was detectable in pods in all stages tested but the third class was consistently the most abundant and more abundant than extensin mRNA (Hong *et al.*,1989). The localisation of these RPRPs demonstrates that they have a different pattern of expression to other cell wall proteins - glycine-rich proteins and extensins (Wyatt *et al.*,1992). Wound induction appears to influence the regulation of one class of RPRP but not another (Klies-San Francisco & Tierney, 1990).

Immunolocalisation studies using antibodies against the second class of RPRP showed it to be localised to the middle lamella and intercellular spaces of the 1-2 day hypocotyl cortex, although extraction studies demonstated that the protein became less soluble and assembled into the cell wall by four days (Marcus *et al.*,1991). This finding is borne out by studies comparing mRNA levels to immunologically detectable protein levels (Klies-San Francisco & Tierney, 1990). Extraction of RPRPs from soybean cell walls reveals that 50% of the proline residues are hydroxylated (Averyhart-Fullard *et al.*,1988, Datta

et al.,1989). The soybean root nodule RPRP appears to be a structural protein involved in nodule morphogenesis rather than in bacterial infection itself (Franssen *et al.*,1987).

Expression of RPRP from other species has not been so well studied. In carrot, RPRP mRNA has been found to increase after wounding but, in contrast to extensin, not as a result of ethylene or fungal elicitor (Chen & Varner, 1985a, Tierney *et al.*,1988). In maize, RPRPs have been extracted from the cell wall fraction in a similar way to extensins and expression is associated with initiation and differentiation of vascular tissue, suggesting a role in the early construction of cell walls. Transient high levels of mRNA are also detected after wounding, peaking at 1-2 hours (Stiefel *et al.*,1988 & 1990).

In summary, it appears that in the species so far studied, RPRPs compose small, differentially expressed, spatially and temporally regulated gene families with some members being wound induced. Expression of the RPRPs is similar to that of extensins with the protein being incorporated into the cell wall of structural cells. The aa composition of the RPRPs with high levels of serine, threonine, hydroxyproline, lysine, tyrosine and histidine allows plenty of scope for intermolecular interactions (Varner & Lin, 1989), although the lack of polyproline peptides possibly endows these proteins with a greater flexibility than extensins.

The C-terminal region of PP590

As previously mentioned the C-terminal region of PP590 was found to have homology to a hydrophobic protein isolated from soybean seeds. This soybean protein is predominantly hydrophobic with a short hydrophilic C-terminal peptide, like the corresponding region of PP590. This protein is present at high levels in the mature soybean seed (200mg/kg) but as

yet, no role has been ascribed to it (Odani *et al.*,1986). Hydrophobic cluster analysis places this protein within a group including lipid transfer proteins and seed storage proteins, sharing structural similarity (Henrissat *et al.*,1988). No DNA sequence data is available for this protein so it is impossible to tell whether it forms part of a larger -possibly proline-rich-protein (like PP590) which is subsequently cleaved.

Another protein with a similar degree of homology (48% identity and 87% homology) to this region of PP590 (fig. 42) is encoded by a gene isolated when an *Arabidopsis* library was screened with an extensin cDNA from *Brassica napus* (Yaish, 1990). This gene has a 381bp ORF encoding 127 residues, comprising a signal peptide, a short (23 residue) basic and proline rich region including a polyproline pentamer (presumably accounting for its hybridisation to the extensin probe), followed by the sequence homologous to PP590 (fig. 42). It is not known whether this gene is expressed, in the tissues tested, the only mRNA hybridised to by this DNA was 1.3kb which seems too large to be encoded by this gene.

The homology between PP590 and the two previously mentioned sequences is, however, completely eclipsed by the homology between it and the C-terminal sequence of a proline-rich protein encoded by a gene expressed in the young tomato fruit, TPRP-F1 (Salts *et al.*,1991 & 1992) (fig. 42). There is 88% identity between the two sequences at the aa. level and 79% between the two nucleotide sequences encoding them. This high level of homology between sequences from two distinct plant families (compared with the 70% identity of the pectinesterase sequences, see below) infers a specific functional role for the whole of this region. Unfortunately, little is known of this tomato sequence other than the pattern of expression of its mRNA, being high in young tomato fruit and low (below 3.5% of this) in stem, root, etiolated

Fig. 42 Alignment of the C-terminal region encoded by PP590 with the C-terminal regions of a tomato proline rich protein (TPRP-F1), a protein encoded for by an *Arabidopsis* gene selected by an extensin probe (*ExtA*) and the sequence of a soybean hydrophobic protein (SHP). : denotes identity and . homology to the PP590 residues, stop codons from the DNA sequences are shown.

```
TPRP-F1  AQPTCPIDALKLGACVDVLGGLIHIGIGGSAKQTCCPLLGGLVDLDAAIC
         :::::.:::::::::::::::::::::::::::::::: :::::::::.:
PP590    AQPTCSIDALKLGACVDVLGGLIHIGIGGSAKQTCCPLLQGLVDLDAAVC
         ..::::. ::::: .:..::. :..... ... .::.:..:::::.:::::
ExtA     PKPTCK-DALKLKVCANVLD-LVKVSL--PPTSNCCALIKGLVDLEAAVC
         ..:.:.    .:. :...::: .      .. ..::.:. :: :..: ::
SHP   ALITRPSCP----DLSICLNILGGSL------GTVDDCCALIGGLGDIEAIVC

      TPRP-F1  -LCTTIRLKLLNINIILPIALQVLIDDCGKYPPKDFKCPST*
               :::::::::::::::..:.:::::::: ::: ::..:::::.
      PP590    -LCTTIRLKLLNINLVIPLALQVLID-CGKTPPEGFKCPSS*
               :::... . :.::: .:..:.:... :::. :.:::::.
      ExtA     -LCTALKANVLGINLNVPISLNVVLNHCGKKVPSGFKCA*
               ::. .:  :  .::   :.: .: . ::.. :.. .::..
      SHP      VLCIQLRA-LGILNLNRNLQL-ILNS-CGRSYPSNATCPRT
```

seedlings, leaf, and mature-green and ripe fruits (Salts *et al.*,1991).

The high level of cysteine and its distribution throughout this region of these proteins is reminiscent of the sequences conserved within families of enzyme inhibitors from plant seeds (Richardson, 1991). Although it is tempting to speculate a similar function could be ascribed to this moiety of PP590, anchored to the pod cell wall by its RPRP region, no further homology to these inhibitors could be found. It is possible that this cysteine distribution denotes a shared structural, rather than functional, similarity.

Hybrid proline-rich proteins

Recently the sequences of a number of "hybrid" proteins have been published. These, like PP590 and the tomato proline-rich protein, not only contain an extensin-like or class IV RPRP-like proline-rich sequence but a second, presumably functional rather than structural

domain. In sunflowers, a family of anther epidermis-specific genes is expressed in late development and accompanied by pigmentation and elongation. These are found to contain short proline-rich regions with the remaining sequence containing a similar high cysteine content to the C-terminal region of PP590 (Domon *et al.*,1990, Evrard *et al.*,1991).

Two wound induced genes from *Phaseolus* fall into this category, one has a C-terminal extensin-like sequence and a 275+ aa N-terminal region of unknown function (Sauer *et al.*,1990), the other resembles a class IV RPRP (with repeats of P-V-H-P-P-V-K-P-P-V and related peptides) but with a non-repetitive C-terminal region low in proline (Sheng *et al.*,1991). Both these genes are present as single copies in the *Phaseolus* genome and are thought to encode cell wall proteins involved in remodelling the plant cell wall during the defence response.

The solanaceous lectins (see above) also belong to this group. It has been suggested, both for these and the *Phaseolus* defence genes, that they have arisen as a result of a gene fusion (Showalter & Varner, 1989). The repetitive nature of the DNA sequence encoding RPRPs must increase the likelihood of recombination occurring and this forms an equally plausable explanation for the origin of PP590.

Comparison of the 5' region of PP590 with that of its tomato counterpart

As stated above, the C-terminal region of PP590 is highly homologous to that of the tomato proline-rich protein, TPRP-F1, and the proline-rich regions also have a similar structure. At first sight the recent publication of the genomic sequence (Salts *et al.*,1992) revealed surprising differences in the 5' region of the coding sequence which had not been covered by the cDNA. An aa sequence is predicted by TPRP-F1 (CPYCPYPPST) resembling the putative N-terminal sequence of PP590

(CPYCPYPSPK), but this begins at residue 7 of the TPRP-F1 sequence compared to residue 25 in PP590. It would appear then that TPRP-F1 lacks a signal peptide, however, when the immediate 5' flanking region of the gene was translated, a reading frame similar to that of the PP590 signal peptide was found:

PP590 M A N F A I A N V L I L L L N L S T L L N V L A:C P

TPRP-F1 M E K F N L A R V L L L L L Q L G T L F I A H A:C P

The reading frame above is not continuous in the published sequence, it includes a frame change and a deletion of four residues from those predicted by the DNA sequence. If the DNA sequence is examined, a direct repeat of 13bp (GCATGTCCTTATT) occurs in the region of the proposed signal peptide cleavage site (: above) of PP590. If this repeat is removed from the TPRP-F1 DNA sequence, the above aa sequence is predicted. These residues constitute a plausable signal peptide with hydrophobic core, LA(R)VLLLLL, and cleavage site (von Heijne, 1983) resulting in the same mature N-terminal peptide (CPCPCY) as that proposed for PP590.

As this gene is otherwise so homologous to PP590 and (for the same reasons as argued for PP590 - see below) is functional, it would appear that this repeat is an artifact. Such an artifact may have arisen by recombination during subcloning or (more likely) during the compilation of the sequence data. No sequences homologous to PP590 could be found in the remaining short (94b) 5' flanking sequence published.

Expression of PP590

Genomic blots indicate that PP590 is present as a single copy in the PP genome and that another single copy gene is present which has homology to it. Like PP590, TPRP-F1 is found as a single copy within the tomato genome (Salts *et al.*,1991). The fact that PP590 is present as a

single copy, together with the identity of the cDNA and genomic sequences, indicates that PP590 encodes the cDNA isolated and is therefore expressed in pods.

A mRNA of 1.75-1.8kb was detected in pod RNAs, this is longer than would be expected with a 1056bp ORF. A hydroxyproline-rich glycoprotein encoding cDNA from *Phaseolus* has a 608bp 3' untranslated region despite containing polyadenylation signals 180 and 320bp distal to the stop codon (Sauer *et al.*,1990). It may be that the first polyadenylation signal at 154bp 3' to the stop codon is disregarded and either that at 685bp, or the multiple overlapping signal at 730bp, are used preferentially, which would account for the size discrepancy. The tomato fruit expressed RPRP gene and two other proline-rich protein genes contain introns in their 3' flanking regions (Chen & Varner, 1985a, Stiefel *et al.*,1990, Salts *et al.*,1992). As the cDNA did not extend into this region, it is not possible to tell whether this is the case in PP590.

When northern blots from other tissues were probed with the cDNA, expression was detected in petals but not leaves. It is interesting that hybridisation occurred to the other pigmented tissue (especially in the context of the discussion below), although further work with non-pigmented petals and other tissues is required before definite conclusions can be drawn. The PP590 gene has tentatively been assigned the name *ptxA* as it appears to be pigmented tissue expressed and of unknown function.

PP590 and the purple podded phenotype

The difference in levels of PP590 mRNA in the pods of the green and purple plants from the PP line is difficult to explain. The sequence bears no homology to any of the enzymes isolated from the anthocyanin

biosynthesis pathway and it seems unlikely that a regulatory protein should contain a putative structural region or that its mRNA should be present at such high levels as are found in the purple line. It may be that transcription of PP590 is controlled by the same factor as that regulating the production of pigment but this begs the question why? Problems were encountered producing complete (non-partial) digests of this region of the genomic DNA. It is possible that methylation or another factor is preventing transcription from this region in the green plants and that PP590 by chance lies near to a gene in, or regulating, the anthocyanin biosynthesis pathway which is also affected. Methylation of DNA has been shown to be a mechanism by which transposable elements are inactivated both in their normal host plants and when inserted into transgenic plants (Fedoroff, 1989, Martin *et al.*,1989, Linn et *al.*,1990). However, two factors argue against methylation being the cause of the poor digestion; firstly, the *Xba* I recognition sequence (TCTAGA) should not be affected by plant methylation (CG and CNG, Gruenbaum *et al.*,1981) and secondly, the plants showing poor digestion are those exhibiting instability and so presumably contain an active unmethylated element. It is not clear whether the "normal" mRNA level is that found in the PP or GP pods. The level in FF pods suggests that the lower level is more usual and the purple to green mutation is causing a return to normal levels found in pods.

Interestingly, in soybean, the expression of a proline-rich protein has also been found to be linked to anthocyanin biosynthesis (Lindstrom & Vodkin, 1991). In this case the level of the protein - the class IV RPRP (SbPRP1) expressed in mature hypocotyl, root and immature seed coat (see above) - was considerably reduced in the mutant pigmented seed coat cultivar when compared to its isogenic cultivar without pigment. mRNA levels mirror the situation at the protein level. These authors

extracted the RPRP using it's affinity to polyvinylpolypyrrolidone (PVPP). However, this method failed to yield any protein, visible on SDS-polyacryamide gel, when attempted on purple pods. The reason for this may be the low tyrosine content of PP590 compared to SbPRP1 (2.8% vs. 16%), as PVPP binds phenolic hydroxyl residues, or differential complexing of phenolic compounds to the proteins, which causes SbPRP1 to bind but not PP590.

The reversal of the relationship of RPRP mRNA level to pigmentation in these two instances lends weight to the argument that a factor is controlling the synthesis of mRNAs encoding both the RPRP and part of the anthocyanin biosynthesis pathway rather than the RPRP itself sequestering precursors from this pathway. A comparison of the flanking regions of the SbPRP1 gene (Hong *et al.*,1987) and PP590 reveals similarly G/C rich sequences in the 5' flanking regions (fig. 32) but no direct homology.

Identification of PP927 as a pectinesterase-like gene

The derived aa sequences of the cDNAs and part of that derived from genomic clone isolated were found, on searching the NBRF protein database, to show homology to tomato pectinesterase. This enzyme has been isolated from tomato fruit and sequenced (Markovic & Jornvall, 1986); it is 305 residues in length predicting a Mr of 33,200. A sequence matching that of the N-terminal region of the mature protein (IIANAVVAQDGTG) occurs in the PP927 derived sequence (IKANVVVAKDGSG) beginning at the 238th residue. The aa sequence C-terminal to this contains regions homologous to peptides sequenced from the tomato enzyme.

The C-terminal region of PP927 appears therefore to encode a pectinesterase. It consists of 317 residues predicting, on its own, a Mr of 34,400. There is close homology with tomato (70% identity to tomato cDNA derived aa sequence, Ray *et al.*,1988) and the aa sequences derived from pectinesterase encoding genes from prokaryotic species are also homologous (20-25% identity) *Erwinia chrysanthemi* (Plastow, 1988), *Aspergillus niger* (Khanh *et al.*,1991) and *Pseudomonas solanacearum* (Spok *et al.*,1991).

Expression of PP927 and other pectinesterase-like genes

The differences found between the cDNA derived from pod tissue and the genomic clone indicate that this gene does not encode this cDNA, and it follows that the gene is not necessarily expressed in pods. The gene on the genomic clone has been designated *pmeA* and that encoding the cDNA *pmeB*. The identity between the cDNA encoded and the genomic clone encoded aa sequences is 80%. This figure is less than that found between the tomato fruit pectinesterase cDNAs and between them and the protein sequence, 87-93% identity (Harriman *et al.*,1991), despite the different

varieties used in each case. This would tend to suggest that pPP927 is expressed elsewhere in the pea plant, although, in tomato, fruit expressed cDNA does not hybridise to mRNA from other tissues showing similar levels of pectinesterase activity (Harriman *et al.*,1991). There is no evidence to suggest that PP927 is not transcribed, it has all the features normally associated with a functional gene, but until an identical cDNA is isolated or S1 nuclease mapping conducted, it is impossible to say that it is functional.

Northern blots with RNA from 5 daf pods contain an mRNA of 2.2kb hybridised to by pPP927 cDNA (fig. 28). This is somewhat longer than would be expected from PP927, with a 1662bp ORF. Possibly the gene which encodes the pod cDNA is longer than PP927, it may have a long 3' untranslated region as is the case with a carrot hudroxyproline-rich glycoprotein (see above) or there may be a discrepancy in the sizing of the bands on gel. Certainly the mRNA is much longer than that predicted by just the pectinesterase region plus flanking sequence and longer than that found in tomato fruit (1.6 kb, Ray *et al.*,1988). There is no difference in mRNA levels between the purple and green podded lines of PP as might be expected, but the FF line contains much less (5%), which accounts for the selection of the cDNA in the differential screen.

When RNAs from different tissues were probed with pPP927 (fig. 30a) no expression could be seen in petal or leaf tissue, certainly not at the levels seen in pod. As with tomato it may be that other tissues express different genes whose products cannot be detected with pod cDNA. More work is required to probe, at varying stringency, RNA from other tissues and those examined so far and to assess mRNA levels in the pod through development.

Western blots appear to indicate the presence of pectinesterase in both pod and seeds from both PP and FF lines, although the result is

rather ambiguous. The blots have high backgrounds and the specificity of these tomato antibodies is unknown in pea. Even within the tomato plant, leaf, stem and root pectinesterase could not be detected with fruit pectinesterase antibodies, although activity could be detected at similar levels (Harriman *et al.*,1991).

Pectinesterase-like genes in the pea genome

The genomic blot (fig. 40) indicates that there are three pectinesterase-like genes in the purple podded line and four in the FF line, with at least two more similar genes present in both genotypes. In tomato a family containing up to three members was identified (Ray *et al.*,1988, Harriman & Handa, 1990) and in *Brassica* at least two members are present (Albani *et al.*,1991). In tomato it has been reported that three pectinesterase genes lie on a single 12kbp genomic clone (Harriman & Handa, 1990). Two of the PP and three of the FF pectinesterase-like sequences lie on identical sized *Eco* RI an *Nsi* I fragments, this suggests that a similar tandem repeat of the genes may occur in pea.

The probe encoding the non pectinesterase-like region of PP927 when washed at high stringency, only detects a single band in each digest. These are the bands predicted by the genomic clone. At low stringency the result is more difficult to interpret, one of other Eco RV fragments hybridised to at high stringency by the 3' probe is detected but neither multiple copy band in the Eco RI and Nsi I digests is detected. This suggests at least one copy of a pectinesterase-like gene in the PP genome, and two in the FF genome, are present without an N-terminal region similar to PP927. These results demonstrate that the N-terminal portion of the PP927 encoded sequence is less well conserved than the pectinesterase region and that some pectinesterase genes may exist without such a region at all. Both probes at low stringency detect the

same series of bands suggesting that there may be a second gene family whose sequences show similarity to both regions of PP927.

Occurrence and function of pectinesterase

Pectinesterase activity has been found in all plants tested for it (Cassab & Varner, 1988). Activity has been demonstrated in pea seeds (Kim & Love, 1990) and *Phaseolus vulgaris* pods (Summers, 1989), hence it would not be surprising to find pectinesterase expressed in pea pods. Pectinesterase is a pectolytic enzyme which catalyses the demethylation of polygalacturonic acid within highly esterified pectin, rendering it susceptible to cleavage by polygalacturonase and pectate lyases (Lee & Macmillan, 1970, Rexova-Benkova & Markovic, 1976). Pectolytic enzymes cause cell wall softening during fruit ripening in tomato, they also play a major role in plant cell wall degredation by pathogenic organisms (Collmer & Keen, 1986).

Pectolytic activity has been found in isolated plant cell walls (Kivilaan *et al.*,1971, Koch & Nevins, 1989) and increased pectinesterase activity has been found to be associated with growth potential in hypocotyl cells of *Vigna radiata* (Goldberg, 1984). It has been suggested that this enzyme plays a crucial part in the process of plant cell wall growth and reorganisation (Yamaoka *et al.*,1983, Varner & Lin, 1989), possibly through its effect on cell wall charge density and Ca^{2+} distribution (Ricard & Noat, 1986, Moustacas *et al.*,1991), although a complete picture of the mechanisms involved has not yet emerged. Pectinesterase acts to de-esterify pectin linearly along the chain (Rexova-Benkova & Markovic, 1976) and de-esterification of blocks of pectin is required before strong gels can be formed in the presence of Ca^{2+} (Rees, 1982). A model for the interaction of de-esterified pectin chains and Ca^{2+} to form an "egg-box" like structure has been proposed

(Grant *et al.*,1973). Acidic, de-esterified pectin molecules have been proposed as a possible site for inter-molecular crosslinkage with the basic lysine residues of extensin molecules (Varner & Lin, 1989), although the distribution of the two components (extensin in the cell wall and pectin in the middle lamella) argues against this (Varner & Hood, 1988).

In situ hybridisation experiments using PP927 as a probe in developing *Arabidopsis* flowers reinforce this role in cell wall modification with pectinesterase mRNA being localised to the region of the gynoecium where the two halves of the septum have just fused (fig. 38). This fusion of the two parts of the septum is followed by the formation, in these regions, of aerenchyma tissue, also associated with high levels of pectinesterase mRNA (Spence,J., personal communication). The middle lamella region between plant cell walls is thought contain high levels of pectins (Darvill *et al.*,1980) and it seems likely that pectinesterase activity would be required in the process "softening" this pectin during aerenchyma cell formation. A build up of pectinesterase mRNA also occurs in the tapetum of the pollen sac prior to its degeneration. Recently, a pectinesterase-like gene expressed in developing pollen from *Brassica napus* has been isolated (Albani *et al.*,1991) and other pectolytic enzyme encoding genes from tomato and *Oenothera* have been found to be pollen-expressed (Wing *et al.*,1990, Brown & Crouch, 1990). In addition, pectolytic enzymes, or their activity, has been detected in pollen from several other plant species including monocots (Pressey & Reger, 1988, Brown & Crouch, 1990), these enzymes may play a role in the growth and penetration of the pollen tube in the gynoecium.

It appears then that rather than just being an enzyme which helps "soften up" ripening fruit, pectinesterase plays an important role in

cell wall growth and reorganisation in the whole plant. The initial action of pectinesterase probably leads to a "firming" of the tissue due to the formation of a more rigid pectin gel structure. The further action of pectinesterase coupled with that of polygalacturonase then leads to tissue softening.

Phenotypic effects of pectinesterase

There seems to be little phenotypic difference betweeen PP and FF pods that can be equated to the differing levels of pectinesterase mRNA. It may be that levels of pectinesterase mRNA increase later in development of the FF pod, hybridisation studies on RNA from both lines of pod through development are required to clarify this. In tomato fruit at least two genes are expressed (Ray *et al.*,1988, Harriman *et al.*,1991), although it seems unlikely that product of a second pod expressed gene would not be detected on the northern blots of FF RNA. Evidence from western blots tends to suggest that at 5 daf, pectinesterase protein levels are unrelated to mRNA levels and that equal amounts of the enzyme are present in both lines. The unknown specificity of the antiserum used and the general high background on these blots makes them rather inconclusive, but no protein was detected in the two PP tracks that was not also detected in the FF track.

Pods from the PP line twist as they desicate, possibly to aid seed dispersal, a feature not found in the FF line, whether this is related to pectin structure is as yet unknown. Premature seed loss is an economically important factor in some crops, such as oil seed rape, and it has been suggested that pectolytic enzymes play a role in this through their action in cell wall degradation. Although raised polygalacturonase levels were not found to be associated with dehiscence (Meakin & Roberts, 1990), the presence of pectinesterase mRNA in the

septum of *Arabidopsis* pods after fertilization may well be worth investigating further in this context.

The level of pectinesterase in pods from different varieties of *Phaseolus vulgaris* has been found to vary by up to four-fold and increased enzyme activity is associated with firm podded varieties (Summers, 1989). Pectinesterase levels have also been investigated in conjunction with seed texture and cooking quality (Jones & Boulter, 1983, Kim & Love, 1990, Bhatty, 1990). Firmness and texture of edible pods and legume seeds is obviously of great importance to the food industry.

Comparison of pectinesterases

Alignment of the derived pectinesterase aa sequences determined so far (fig. 43) confirms the presence of highly conserved regions between and within the eukaryotic and prokaryotic sequences (Albani *et al.*,1991, Spok *et al.*,1991). These regions presumably form part of the functional site of the pectinesterase enzyme. The position of these regions within the sequence does not correlate with regions of hydrophilicity or hydrophobicity. Similarity has been shown between the aa sequence of the pectinesterase gene from *Erwinia chrysanthemi* and that of pectate lyase genes from *Erwinia* species and tomato (Wing *et al.*,1990). However, the conserved regions of the pectate lyase genes are not those conserved between the pectinesterase genes.

The N-terminal region of PP927

As previously stated, the protein sequence of this gene has homology to cDNAs encoding tomato pectinesterase. Homology extends beyond the N-terminal aa of the mature protein in tomato (Markovic and Jornvall, 1986) and beyond the previously assigned start codon, to the

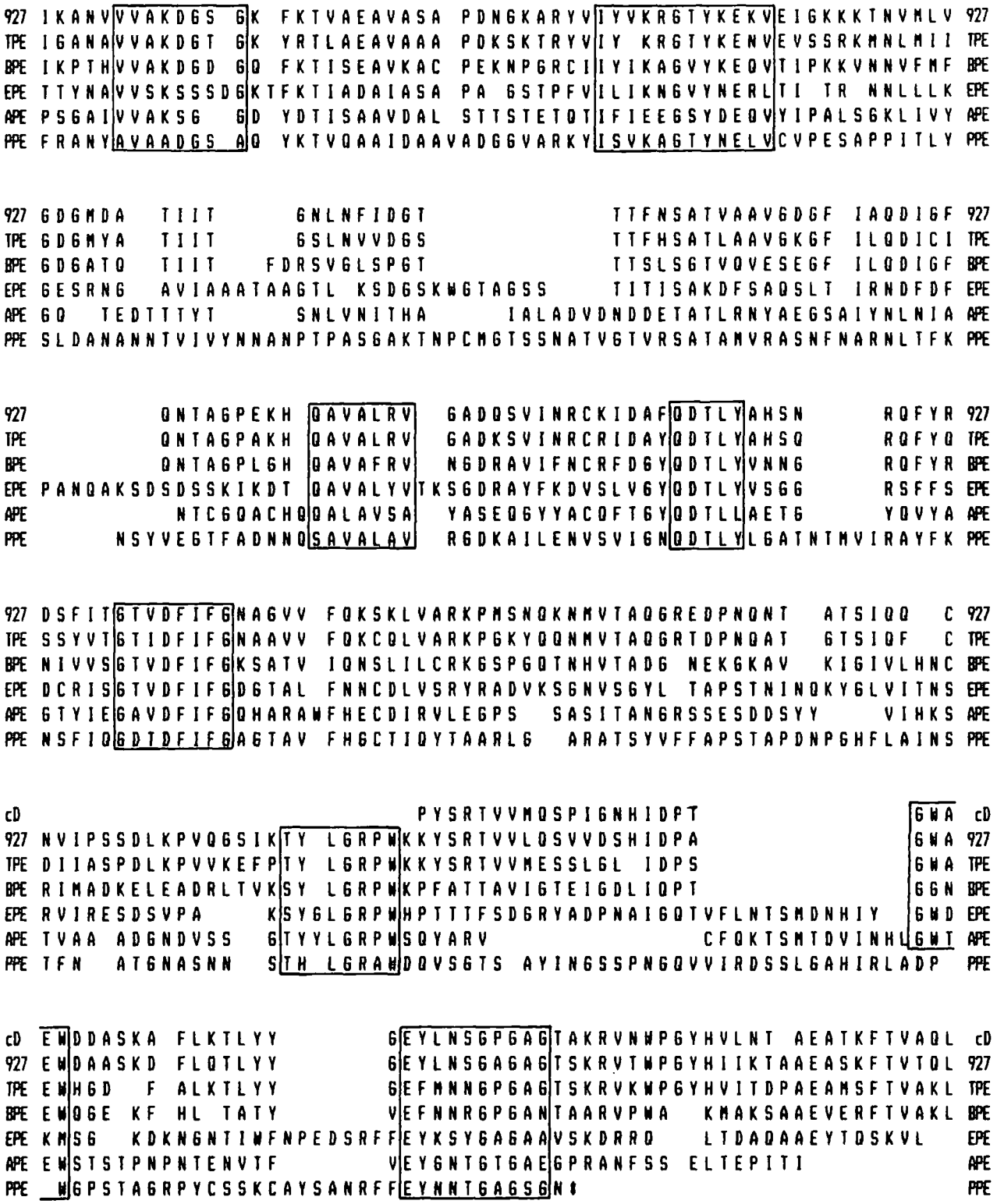
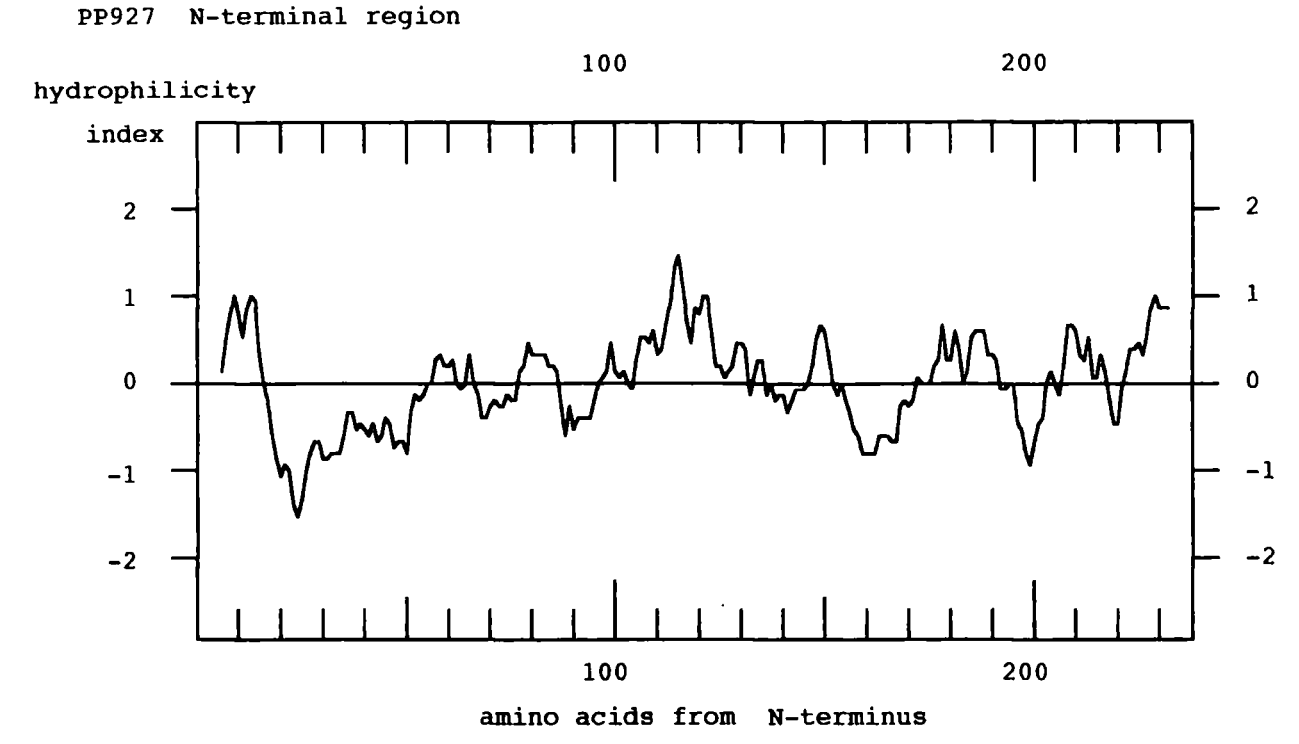


Fig. 43 Comparison of deduced amino acid sequences of pectinesterase genes (or pectinesterase-like regions) from different species. Genes abbreviated as follows: cD, pPPL927; 927, PP927; TPE, tomato pectinesterase; BPE, *Brassica* pectinesterase; EPE, *Erwinia* pectinesterase; APE, *Aspergillus* pectinesterase; PPE, *Pseudomonas* pectinesterase. Conserved regions are boxed.

end of the cDNA which extends further in this region, pPE1 (Ray *et al.*,1988). This indicates that pPE1 is in fact, not a full length cDNA. There are 237 residues C-terminal from the start of the PP927 ORF to the N-terminus of the mature pectinesterase protein from tomato, using this region to search the NBRF protein database yielded no homologous proteins. The sequence has 45% identity with the 110aa encoded by the tomato cDNA 5' region and optimised alignment of the PP927 N-terminal region with the same region encoded by the *Brassica napus* pectinesterase-like gene revealed 15% identity and 61% homology (when conservative substitutions were included). This shows a significant match (8.5 standard deviations above the mean for randomised sequence) using the FASTA package (Pearson & Lipman, 1988). Hydrophilicity plots (Hopp & Woods, 1981) for these two regions (fig. 44) show they share a similar composition but no outstanding features.

There appears to be a signal peptide at the N-terminus with a hydrophobic core FWLIL(S)LAAII and a predicted cleavage site (von Heijne, 1983) between the 33rd and 34th residues. The N-terminal sequence of the whole protein would then be LIVSHLNKPI and its Mr 56,800, 34,400 accounted for by the pectinesterase-like region and 22,400 by the N-terminal region.

The function of the N-terminal region encoded by PP927 is unclear (as is that for this region encoded by the *Brassica* gene); it may be that this region functions to locate the enzyme - presumably in the cell wall, or it may act to inactivate the enzyme until its activity is required and cleavage occurs. The levels in this region of the aa (serine 13%, threonine 8%, proline 3%, lysine 5%, tyrosine 0, & histidine 3.5%) usually associated with cell wall interactions (Varner & Lin, 1989) do not strongly infer a role in situating this protein within the cell wall.



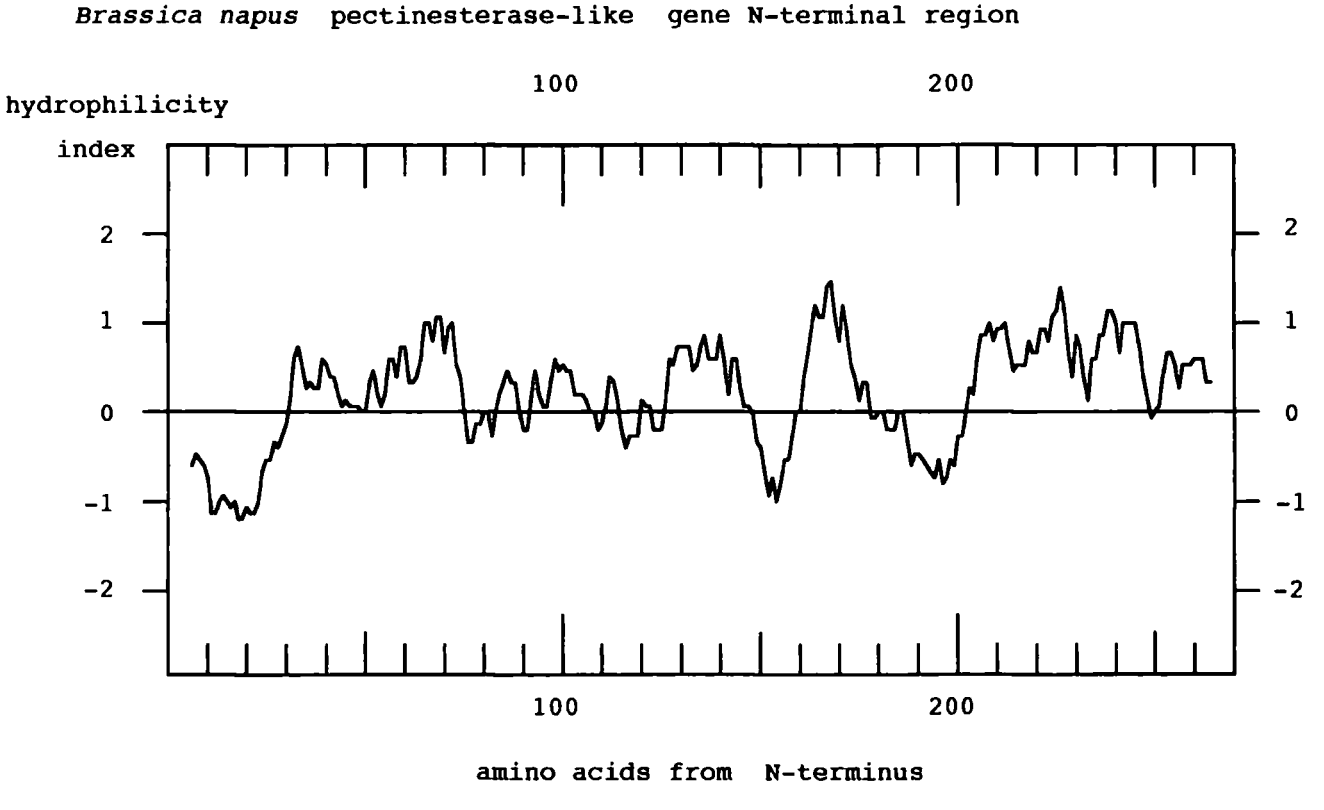


Fig. 44 Hopp Woods hydrophilicity plots for N-terminal regions (N-terminal to thr pectinesterase-like regions) encoded by PP927 (above) and the *Brassica napus* pectinesterase-like gene (below).

Whatever the function of this N-terminal region, genomic blots show that it is not highly conserved between the pectinesterase-like genes in pea. This supports a role requiring a similar overall aa composition or final structure, rather than a conserved active site such as that in the pectinesterase-like region. The Mrs of proteins detected on western blot (28k pod & 34k seed) suggest that pectinesterases present in the 5 daf pod and mature seed do not have N-terminal extensions, although it is not possible to tell whether PP927 is expressed in these organs. It may be that the presence of the N-terminal region would inhibit antiboby from binding to the pectinesterase region.

CHAPTER FOURTEEN: CONCLUDING DISCUSSION

Two of the three storage protein genes initially sequenced proved to be pseudogenes, this, together with the results from other similar genes, demonstrates a fairly high level of redundancy amongst pea storage protein genes. The line to line variability in the composition of the storage protein component of pea seeds suggests that a wide variability can be tolerated and that the loss of function of individual genes is not catastrophic. Comparison of the sequences from the convicilin and legumin gene families suggest that the addition and rearrangement of acidic hydrophilic regions is a mechanism for mutation of these genes which has occurred a number of times in evolution.

The 5' flanking regions of these genes have been compared with those of other vicilin and legumin encoding genes and conserved sequences identified. Whilst this has not revealed any major new consensus sequences, this work confirms previous findings and adds to the data which may be used to assess the results obtained by DNA-protein binding assays and plant transformation with deletion/substitution mutants. Two results are particularly interesting in this context; firstly the finding that sequences in the 5' flanking region of *cvcA* compete for binding to a nuclear protein which binds to the 5' flanking region of *legA*, and secondly, that despite the similarity in the pattern of expression in the developing seed between *cvcA* and *legK*, no specific sequences with significant homology could be found between their 5' flanking regions.

The specific gene product of *cvcA* was determined by transfer of the gene to transgenic tobacco, it was a minor component of convicilin with a lower Mr than the more abundant species. The presence of a functional *legK* gene in pea line Birte was confirmed and with this information the product of this gene was identified by its absence from the protein

extract of the line with the pseudogene. Identification of this gene's product in conjunction with other work has enabled further elucidation of the legumin phenotype. *vicJ* appears to be a pseudogene in the lines investigated due to sequence rearrangement at the 3' end of the coding sequence and no evidence for its expression could be found.

The isolation of the major root protein was successful although the sequence data obtained from it could not be matched to any known sequences. Further work is necessary to isolate more protein and obtain further sequence from internal tryptic peptides. A total amino acid composition might help to classify the protein itself. Extra internal aa sequence could be used to synthesise another oligonucleotide, with the aim of using PCR to amplify sequence encoding this protein which could then be used as a probe for the whole gene.

The search for the cause of the instability in the purple podded phenotype has so far been unsuccessful. Similar levels of mRNA were present in both the mutant green and purple podded lines when probed with sequences encoding chalcone synthase, chalcone isomerase and dihydroflavonol reductase enzymes from the anthocyanin biosynthesis pathway. Two of the cDNAs isolated by the differential screen show reduced expression in the mutant green podded line compared to the purple line but whether they are involved in the production of pigment is unclear. It cannot be determined from the results obtained so far whether these sequences are affected directly by the presence of a transposable element or whether their expression is regulated by a gene product which is. There does not appear to be any polymorphism in the sequence adjacent to the PP590 (*ptxA*) gene in the pea plants exhibiting instability, although the problems with cleavage of the DNA in this region may indicate that it is the object of selective methylation.

The two genes isolated as a result of the differential screen both

proved to be interesting in themselves. The pectinesterase-like gene is present as part of a small multigene family in pea and is not identical to the cDNA from pod used to isolate it. The role of pectinesterase in plant cell wall rearrangement is not clear at present. The increased expression in the Purple Podded line compared with the Feltham First line suggests that it has a specific role in PP pods in addition to its role in cell growth. The tissue specific expression of the pectinesterase-like sequences and the PP590 sequence is at present under investigation but initial results link PP590 to pigmented tissue and show high level expression of the pectinesterase genes to be pod specific. The proline-rich nature of the N-terminal region of the PP590 sequence suggests that it is located in the cell wall but the function of the C-terminal region is unclear. The isolation of these two pod genes, both with counterparts expressed in the young tomato, suggests that despite their divergent morphologies, similar patterns of selective gene expression occur in these two fruit.

LITERATURE CITED

Albani,D., Altosaar,I., Arnison,PG. and Fabijanski,SF. (1991). A gene showing sequence similarity to pectin esterase is specifically expressed in developing pollen of *Brassica napus*. Sequences in its 5' flanking region are conserved in other pollen-specific promoters. Plant Mol. Biol. *16*, 501-513.

Allen,G. (1981). Sequencing of proteins and peptides. In, Laboratory techniques in biochemistry & molecular biology. (Work,TS. and Burdon,RH. eds.), vol.9. Elsevier/North Holland, Amsterdam.

Averyhart-Fullard,V., Datta,K. and Marcus,A. (1988). A hydroxyproline-rich protein in the soybean cell wall. Proc. Natl. Acad. Sci. USA *85*, 1082-1085.

Baumlein,H., Wobus,U., Pustell,J. and Kafatos,FC. (1986). The legumin gene family: structure of a B type gene of *Vicia faba* and a possible legumin gene specific regulatory element. Nucl. Acids Res. *14*, 2707-2720.

Bhattacharyya,MK., Smith,AM., Ellis,THN., Hedley,C. and Martin,C. (1990). The wrinkled-seed character of pea described by Mendel is caused by a transposon-like insertion in a gene encoding starch-branching enzyme. Cell *60*, 115-122.

Bhatty,RS. (1990). Cooking quality of lentils: the role of structure and composition of cell walls. J. Agric. Food Chem. *38*, 376-383.

Birnboim,HC. and Doly,J. (1979). A rapid alkaline extraction proceedure for screening recombinant plasmid DNA. Nucl. Acids Res. *7*, 1513-1523.

Blixt,S. (1962). Studies in induced mutations in peas. VI. Mutations in seed-colour, flower-colour, maculum-colour, pod-colour and grey spotting of leaves. Agri. Hort. Genet. *20*, 95-110.

Blixt,S. (1976). The pea. In, Handbook of genetics, pp181-221. (King,RC. ed.) Plenum, New York.

Bonas,U., Sommer,H. and Saedler,H. (1984). The 17-kb *Tam1* element of *Antirrhinum majus* induces a 3bp duplication upon integration into the chalcone synthase gene. EMBO J. *3*, 1015-1019.

Boulter,D., Evans,IM., Ellis,JR., Shirsat,AH., Gatehouse,JA. and Croy,RRD. (1987). Differential gene expression in the development of *Pisum sativum*. Plant Physiol. Biochem. *25*, 283-289.

Boulter,D., Croy,RRD., Evans,IM., Gatehouse,JA., Harris,N., Shirsat,A. and Thompson,A. (1990). The molecular biology of pea seed developement with particular reference to the storage protein genes. in, Genetic engineering of crop plants. (Lycett,GW. and Grierson,D. eds.) Butterworths, London

Bown,D., Levasseur,M., Croy,RRD., Boulter,D. and Gatehouse,JA. (1985). Sequence of a pseudogene in the legumin gene family of pea (*Pisum sativum* L.). Nucl. Acids Res. *13*, 4527-4538.

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

Breathnach,R., Benoist,C., O'Hare,K., Gannon,F. and Chambon,P. (1978). Ovalbumin gene: Evidence for a leader sequence in mRNA & DNA sequence at the exon-intron boundaries. Proc. Natl. Acad. Sci. USA *75*, 4853-4857.

Brown,SM. and Crouch,ML. (1990). Characterization of a gene family abundantly expressed in *Oenothera oreganensis* pollen that shows sequence similarity to polygalacturonase. Plant Cell *2*, 263-274.

Bustos,MM., Guiltinan,MJ., Jordano,J., Begum,D., Kalkan,FA. and Hall,TC. (1989). Regulation of β-glucuronidase expression in transgenic tobacco plants by an A/T-rich *cis*-acting sequence found upstream of a french bean β-phaseolin gene. Plant Cell *1*, 839-853.

Caelles,C., Delseny,M. and Puigdomenech,P. (1992). The hydroxyproline-rich glycoprotein gene from *Oryza sativa*. Plant Mol. Biol. *18*, 617-619.

Casey,R. (1979). Genetic variability in the structure of the α subunits of legumin from *Pisum* - a two dimensional gel electrophoretic study. Heredity *43*, 265-272.

Casey,R., March,JF. and Sanger,E. (1981). N-terminal amino acid sequence of β-subunits of legumin from *Pisum sativum*. Phytochemistry *20*, 161-163.

Casey,R., March,JF., Sharman,JE. and Short,MN. (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.). Biochem. Biophys. Acta *670*, 428-432.

Casey,R., Domoney,C. and Stanley,J. (1984). Convicilin mRNA from Pea (*P. sativum* L.) has sequence homology with other legume 7S storage protein mRNA. Biochem. J. *224*, 661-666.

Cassab,GI. and Varner,JE. (1988). Cell wall proteins. Annu. Rev. Plant Physiol., Plant Mol. Biol. *39*, 321-353.

Chandler,PM., Spencer,D., Randall,PJ. and Higgins,TJV. (1984). Influence of sulfur nutrition on developmental patterns of some major pea seed proteins and their mRNAs. Physiol. Plant. *75*, 651-657.

Chang,JY., Brauer,D. and Wittmann-Liebold,B. (1978). Micro-sequence analysis of peptides and proteins using 4-NN-Dimethylamino-azobenzene 4'-isothiocyanate/phenylisothiocyanate double coupling method. FEBS Lett. *93*, 205-214.

Chen,J. and Varner,JE. (1985a). An extracellular matrix protein in plants: characterization of a genomic clone for carrot extensin. EMBO J. *4*, 2145-2151.

Chen,J. and Varner,JE. (1985b). Isolation and characterisation of cDNA clones for carrot extensin and a proline-rich 33-kDa protein. Proc. Natl. Acad. Sci. USA *82*, 4399-4403.

Chen,Z-L., Schuler,MA. and Beachy,RN. (1986). Functional analysis of regulatory elements in a plant embryo-specific gene. Proc. Natl. Acad. Sci. USA *83*, 8560-8563.

Chen,Z-L., Pan,N-S. and Beachy,RN. (1988). A DNA sequence element that confers seed-specific enhancement to a constitutive promoter. EMBO J. *7*, 297-302.

Chlan,CA., Pyle,JB., Legocki,AB. and Dure,L. (1986). Developmental biochemistry of cotton seed embryogenesis and germination XVIII cDNA and amino acid sequences of members of the storage protein families. Plant Mol. Biol. *7*, 475-489.

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

Clarke,L. and Carbon,J. (1976). A colony bank containing synthetic ColE1 hybrid plasmids representitive of the entire *E. coli* genome. Cell *9*, 91.

Coen,ES. and Carpenter,R. (1986). Transposable elements in *Antirrhinum majus*: generators of genetic diversity. Trends Genet. *2*, 292-296.

Coen,ES. and Carpenter,R. (1988). A semi dominant allele *niv 5-525* acts *in trans* to inhibit expression of its wild type homologue in *Antirrhinum majus*. EMBO J. *7*, 877-883.

Coen,ES., Almeida,J., Robbins,TP., Hudson,A. and Carpenter,R. (1988). Molecular analysis of genes determining spatial patterns in *Antirrhinum majus*. In, Temporal and spacial regulation of plant genes. (Verma,DPS. and Goldberg,RB. eds.), pp63-82. Springer-Verlag, Vienna.

Collmer,A. and Keen,NT. (1986). The role of pectic enzymes in plant pathogenesis. Annu. Rev. Phytopath. *24*, 383-409.

Cone,KC., Burr,FA. and Burr,B. (1986). Molecular analysis of the maize anthocyanin regulatory locus *C1*. Proc. Natl. Acad. Sci. USA *83*, 9631-9635.

Croy,RRD. and Gatehouse,JA. (1985). Genetic engineering of seed proteins; current and potential applications. In, Plant genetic engineering. (Dodds,JH. ed.), pp143-268. Cambridge University Press, Cambridge.

Croy,RRD., Derbyshire,E., Krishna,TG. and Boulter,D. (1979). Legumin of *Pisum sativum* and *Vicia faba*. New Phytol. *83*, 29-35.

Croy,RRD., Gatehouse,JA., Tyler,M. and Boulter,D. (1980a). The purification and characterisation of a third storage protein (convicilin) from the seeds of pea (*P. sativum* L.). Biochem. J. *191*, 509-516.

Croy,RRD., Gatehouse,JA., Evans,IM. and Boulter,D. (1980b). Characterization of the storage protein subunits synthesised *in vitro* by polyribosomes and RNA from developing seeds of pea (*Pisum sativum* L.). I. Legumin. Planta *148*, 49-56.

Croy,RRD., Lycett,GW., Gatehouse,JA., Yarwood,JN. and Boulter,D. (1982). Cloning and analysis of cDNAs encoding plant storage protein precursors. Nature *285*, 76-79.

Croy,RRD., Evans,IM., Yarwood,JN., Harris,N., Gatehouse,JA., Shirsat,AH., Kang,A., Ellis,JR., Thompson,AJ. and Boulter,D. (1988). Expression of pea legumin sequences in pea, *Nicotiana* and yeast. Biochem. Physiol. Pflanzen *183*, 183-197.

Darvill,A., McNeil,M., Albersheim,P. and Delmer,DP. (1980). The primary cell walls of flowering plants. In, The Biochemistry of plants. (Stumpf,PK. and Conn,EE. eds.), vol 1, pp91-162. Academic Press, New York.

Datta,K., Schmidt,A. and Marcus,A. (1989). Characterization of two soybean repetitive proline rich proteins and a cognate cDNA from germinated axes. Plant Cell *1*, 945-952.

Davey,RA. and Dudman,WF. (1979). The carbohydrate of storage glycoproteins from seeds of *Pisum sativum*; characterization and distribution of component polypeptides. Aust. J. Plant Physiol. *6*, 435-447.

De Haan,H. (1930). Contributions to the genetics of *Pisum*. Genetica *12*, 321-439.

Derbyshire,E., Wright,DJ. and Boulter,D. (1976). Legumin and vicilin storage proteins of legume seeds. Phytochemistry *15*, 3-24.

Dickinson,CD., Evans,RP. and Nielsen,NC. (1988). RY repeats are conserved in the 5'-flanking regions of legume seed-protein genes. Nucl. Acids Res. *16*, 371.

Dierks,P., Van Ooyen,A., Cochran,MD., Dobkin,C., Reiser,J. and Weissmann,C. (1983). Three regions upstream from the cap site are required for efficient and accurate transcription of the rabbit β-globin gene in mouse 3T6 cells. Cell *32*, 695-706.

Domon,C., Evrard,J-L., Herdenberger,F. and Pillay,DTN. (1990). Nucleotide sequence of two anther-specific cDNAs from sunflower (*Helianthus annuus* L.). Plant Mol. Biol. *15*, 643-646.

Domoney,C. and Casey,R. (1983). Cloning and characterisation of cDNA for convicilin a major seed storage protein in *Pisum sativum* L. Planta *159*, 446-453.

Domoney,C. and Casey,R. (1984). Storage protein precursor polypeptides in cotyledons of *Pisum sativum* L. Identification of and isolation of a cDNA clone for an 80,000 Mr legumin related polypeptide. Eur. J. Biochem. *139*, 321-327.

Domoney,C. and Casey,R. (1985). Measurement of gene number for seed storage proteins in *Pisum*. Nucl. Acids Res. *13*, 687-699.

Domoney,C. and Casey,R. (1987). Changes in legumin mRNAs throughout developement in *Pisum sativum* L. Planta *170*, 562-566.

Domoney,C. and Casey,R. (1990). Another class of vicilin gene in *Pisum*. Planta *182*, 39-42.

Domoney,C., Ellis,THN. and Davies,DR. (1986a). Organization and mapping of legumin genes in *Pisum*. Mol. Gen. Genet. *202*, 280-285.

Domoney,C., Barker,D. and Casey,R. (1986b). The complete deduced amino acid sequence of legumin β-polypeptides from different genetic loci in *Pisum*. Plant Mol. Biol. *7*, 467-474.

Dooner,HK. (1983). Coordinate genetic regulation of flavonoid biosynthetic enzymes in maize. Mol. Gen. Genet. *189*, 136-141.

Dooner,HK., Robbins,TP. and Jorgensen,RA. (1991). Genetic and developmental control of anthocyanin biosynthesis. Annu. Rev. Genet. *25*, 173-199.

Doring,HP., Tillmann,E. and Starlinger,P. (1984). DNA sequence of the maize transposable element *Dissociation*. Nature *307*, 127-130.

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

Dynan,WS. and Tjian,R. (1985). Control of eukaryotic messenger RNA synthesis by sequence specific DNA-binding proteins. Nature *315*, 774-778.

Ebel,J. and Hahlbrock,K. (1982). Biosynthesis. In, The flavonoids: Advances in research. (Harborne,JB. and Mabry,TJ. eds.), pp641-679. Chapman & Hall, London.

Edwards,GA. (1988). Plant transformation using an *Agrobacterium tumefaciens* Ti-plasmid vector system. Ph.D. Thesis, Durham University.

Edwards,JW. and Coruzzi,GM. (1990). Cell specific gene expression in plants. Annu. Rev. Genet. *24*, 275-303.

Ellis,THN., Davies,DR., Castleton,JA. and Bedford,ID. (1984). The organisation and genetics of rDNA length variants in peas. Chromosoma *91*, 74-81.

Ellis,THN., Domoney,C., Castleton,J., Cleary,W. and Davies,DR. (1986). Vicilin genes of *Pisum*. Mol. Gen. Genet. *205*, 164-169.

Epstein,L and Lamport,DTA. (1984). An intramolecular linkage involving isodityrosine in extensin. Phytochemistry *23*, 1241-1246.

Evans,IM., Gatehouse,JA., Croy,RRD. and Boulter,D. (1984). Regulation of the transcription of storage-protein mRNA in nuclei isolated from developing pea (*Pisum sativum* L.). Planta *160*, 559-568.

Evans,IM., Gatehouse,LN., Gatehouse,JA., Yarwood,JN., Boulter,D. and Croy,RRD. (1990). The extensin gene family in oilseed rape (*Brassica napus* L.): Characterisation of sequences of representitive members of the family. Mol. Gen. Genet. *223*, 273-287.

Evrard,J-L., Jako,C., Saint-Guily,A., Weil,J-H. and Kuntz,M. (1991). Anther specific, developmentally regulated expression of genes encoding a new class of proline-rich proteins in sunflower. Plant Mol. Biol. *16*, 271-281.

Favaloro,J., Treisman,R. and Kamen,R. (1980). Transcription maps of polyoma virus-specific RNA: analysis by two-dimentional nuclease S1 gel mapping. Methods Enzymol. *65*, 718-749.

Fedoroff,NV. (1989). Maize transposable elements. In, Mobile DNA. (Berg,DE. and Howe,MM. eds.), pp375-411. Am. Soc. Microbiol., Washington DC.

Fedoroff,NV., Mauvais,J. and Chaleff,D. (1983). Molecular studies on mutations at the *Shrunken* locus in maize caused by the controlling element *Ds*. J. Mol. Appl. Genet. *2*, 11-29.

Fedoroff,NV., Furtek,DB. and Nelson,OE.Jr. (1984). Cloning of the *bronze* locus in maize by a simple and generalizable procedure using the transposable controlling element *Activator (Ac)*. Proc. Natl. Acad. Sci. USA *81*, 3825-3829.

Feinberg,AP. and Vogelstein,B. (1984). Addendum to: A technique for radiolabelling DNA restriction endonuclease fragments to high specific activity. Anal. Biochem. *137*, 266-267.

Finnegan,DJ. (1989). Eukaryotic transposable elements and genome evolution. Trends Genet. *5*, 103-107.

Franssen,HJ., Hap,J-P., Gloudemans,T., Stiekema,W., van Dan,H., Govers,F., Louwerse,J., van Kammen,A and Bisseling,T. (1987). Characterization of a cDNA for nodulin-75 of soybean: A gene product involved in early stages of root nodule development. Proc. Natl. Acad. Sci. USA *84*, 4495-4499.

Freeling,M (1984). Plant transposable elements and insertion sequences. Annu. Rev. Plant Physiol. *35*, 277-298.

Gatehouse,JA. and Boulter,D. (1980). Isolation and properties of a lectin from the roots of *Pisum sativum* (Garden pea). Physiol. Plant. *49*, 437-442.

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

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

Gatehouse,JA., Lycett,GW., Croy,RRD. and Boulter,D. (1982a). The post translalional proteolysis of the subunits of vicilin from pea (*Pisum sativum* L.). Biochem. J. *207*, 629-632.

Gatehouse,JA., Evans,IM., Bown,D., Croy,RRD. and Boulter,D. (1982b). Control of storage-protein synthesis during seed development in pea (*Pisum sativum* L.). Biochem. J. *208*, 119-127.

Gatehouse,JA., Lycett,GW., Delauney,AJ., Croy,RRD. and Boulter,D. (1983). Sequence specificity of the post-translational proteolytic cleavage of vicilin, a seed storage protein of pea (*Pisum sativum* L.). Biochem. J. *212*, 427-432.

Gatehouse,JA., Croy,RRD. and Boulter,D. (1984). The synthesis and structure of pea seed storage proteins. In, Critical reviews in plant sciences, vol.1, pp287-314. C.R.C. Press, Boca Raton, Fla.

Gatehouse,JA., Evans,IM., Croy,RRD. and Boulter,D. (1986). Differential expression of genes during legume seed developement. Phil. Trans. R. Soc. Lond. *B314*, 367-384.

Gatehouse,JA., Bown,D., Evans,IM., Gatehouse,LN., Jobes,D, Preston,P. and Croy,RRD. (1987). Sequence of the seed lectin gene from pea (*Pisum sativum* L.). Nucl. Acids Res. *15*, 7642.

Gatehouse,JA., Bown,D., Gilroy,J., Levasseur,M., Castleton,J. and Ellis,THN. (1988). Two genes encoding "minor" legumin polypeptides in pea (*Pisum sativum* L.). Biochem. J. *250*, 15-24.

Gatehouse,LN. (1985). Construction of a cDNA library encoding pea seed proteins. MSc. Thesis, Durham University.

Gatehouse,LN., Evans,IM., Gatehouse,JA. and Croy,RRD. (1990). Characterisation of a rape (*Brassica napus* L.) extensin gene encoding a polypeptide relatively rich in tyrosine. Plant Sci. *71*, 223-231.

Goldberg,R. (1984). Changes in the properties of cell wall pectin methylesterase along the *Vigna radiata* hypocotyl. Physiol. Plant. *61*, 58-63.

Goldberg,RB. (1986). Regulation of plant gene expression. Phil. Trans. R. Soc. Lond. *B314*, 343-353.

Goodrich,J., Carpenter,R. and Coen,ES. (1992). A common gene regulates pigmentation pattern in diverse plant species. Cell *68*, 955-964.

Graham,DE. (1978). The isolation of high molecular weight DNA from whole organisms or large tissue masses. Anal. Biochem. *85*, 609-613.

Grant,GT., Morris,ER., Rees,DA., Smith,PTC. and Thom,D. (1973). Biological interactions between polysaccharides and divalent cations: The egg-box model. FEBS Lett. *32*, 195-198.

Grosveld,GC., Shewmaker,CK., Jat,P. and Flavell,RA. (1981). Localization of DNA sequences necessary for transcription of the rabbit β-globin gene *in vitro*. Cell *25*, 215-226.

Gruenbaum,Y., Naveh-Many,T., Cedar,H. and Razin,A. (1981). Sequence specificity of methylation in higher plants. Nature *292*, 860-862.

Hall,TC., Ma,Y., Buchbinder,BU., Pyne,JW., Sun,SM. and Bliss,FA. (1978). Messenger RNA for G1 protein of French bean seeds: Cell-free translation and product characterization. Proc. Natl. Acad. Sci. USA *75*, 3196-3200.

Hames,BD. (1981). An introduction to polyacrylamide gel electrophoresis. In, Gel electrophoresis of proteins: a practical approach. (Hames,BD. and Rickwood,D. eds.), pp1-91. IRL Press, Oxford.

Harker,CL., Ellis,THN. and Coen,ES. (1990). Identification and genetic regulation of the chalcone synthase multigene family in pea. Plant Cell *2*, 185-194.

Harriman,RW. and Handa,AK. (1990). Identification and characterization of three pectin methylesterase genes in tomato (abstract No.249). Plant Physiol. *93S*, 44.

Harriman,RW., Tieman,DM. and Handa,AK. (1991). Molecular cloning of tomato pectin metylesterase gene and its expression in Rutgers, ripening inhibitor, non-ripening and never ripe tomato fruits. Plant Physiol. *97*, 80-87.

Heller,W. (1986). Flavonoid biosynthesis, an overview. In, Plant flavonoids in biology and medicine. (Cody,V., Middleton,E.Jr. and Harborne,JB. eds.), pp25-42. AR. Liss, New York.

Heller,W. and Forkmann,G. (1988). Biosynthesis. In, The Flavonoids: Advances in research since 1980. (Harborne,JB. ed.), pp399-425. Chapman & Hall, London.

Henrissat,B., Popineau,Y. and Kader,J-C. (1988). Hydrophobic cluster analysis of plant protein sequences. Biochem. J. *255*, 901-905.

Higgins,TJV. and Spencer,D. (1981). Precursor forms of pea vicilin subunits. Plant Physiol. *67*, 205-211.

Higgins,TJV., Newbigin,EJ., Spencer,D., Llewellyn,DJ. and Craig,S. (1988). The sequence of a pea vicilin gene and its expression in transgenic tobacco plants. Plant Mol. Biol. *11*, 683-695.

Hirano,H., Gatehouse,JA. and 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.

Hong,JC., Nagao,RT. and Key,JL. (1987). Characterization and sequence analysis of a developmentally regulated putative cell wall protein gene isolated from soybean. J. Biol. Chem. *262*, 8367-8376.

Hong,JC., Nagao,RT. and Key,JL. (1989). Developmentally regulated expression of soybean proline-rich cell wall protein genes. Plant Cell *1*, 937-943.

Hopp,TP. and Woods,KR. (1981). Prediction of protein antigenic determinants from amino acid sequences. Proc. Natl. Acad. Sci. USA *78*, 3824-3828.

Ichinose,Y., Kawamata,S., Yamada,T., An,CC., Kajiwara,T., Shiraishi,T. and Oku,H. (1992). Molecular cloning of chalcone synthase cDNAs from *Pisum sativum*. Plant Mol. Biol. *18*, 1009-1012.

Jofuku,KD., Schipper,RD. and Goldberg,RB. (1989). A frame-shift mutation prevents Kunitz trypsin inhibitor mRNA accumulation in soybean embryos. Plant Cell *1*, 427-435.

Jones,PMB. and Boulter,D. (1983). The cause of reduced cooking rate in *Phaseolus vulgaris* following adverse storage conditions. J. Food Sci. *48*, 623-626.

Jordano,J., Almoguera,C. and Thomas,TL. (1989). A sunflower helianthinin gene upstream sequence ensemble contains an enhancer and sites of nuclear protein interaction. Plant Cell *1*, 855-866.

Joshi,CP. (1987). An inspection of the domain between putative TATA box and translation start site in 79 plant genes. Nucl. Acids Res. *15*, 6643-6648.

Karlovsky,P. (1990). Missuse of PCR. Trends Biol. Sci. *15*, 419.

Khanh,NQ., Ruttkowski,E., Leidinger,K., Albrecht,H. and Gottschalk,M. (1991). Characterisation and expression of a genomic pectin methylesterase-encoding gene in *Aspergillus niger*. Gene *106*, 71-77.

Kieliszewski,M. and Lamport,DTA. (1986). Cross-reactivities of polyclonal antibodies against extensin precursors determined via ELISA techniques. Phytochemistry *25*, 673-677.

Kieliszewski,M. and Lamport,DTA. (1988). Tying the knots in the extensin network. In, Self assembling architecture. (Varner, JE. ed.), pp61-76. AR.Liss, New York.

Kieliszewski,M., de Zacks,R., Leykam,JF. and Lamport,DTA. (1992). A repetitive proline-rich protein from the gymnosperm Douglas fir is a hydroxyproline-rich glycoprotein. Plant Physiol. *98*, 919-926.

Kim,Y-H. and Love,MH. (1990). Pectin methyl esterase activity changes during the developement of the green pea (*Pisum sativum*) to harvest maturity. Abs. Papers Am. Chem. Soc. *200*, 3-4.

Kivilaan,A., Bandurski,RS. and Schulze,A. (1971). A partial characterization of an autolytically solubilized cell wall glucan. Plant Physiol. *48*, 389-393.

Klies-San Francisco,SM. and Tierney,ML. (1990). Isolation and characterization of a proline-rich cell wall protein from soybean seedlings. Plant Physiol. *94*, 1897-1902.

Koch,JL. and Nevins,DM. (1989). Tomato fruit cell wall: 1. Use of purified tomato polygalacturonase and pectinmethylesterase to identify developmental changes in pectins. Plant Physiol. *91*, 816-822.

Koes,RE., Spelt,CE., van den Elzen,PJM. and Mol,JNM. (1989). Cloning and molecular characterization of the chalcone synthase multigene family of *Petunia hybrida*. Gene *81*, 245-257.

Konigsberge,W. (1972). Reduction of disulphide bonds in proteins with dithiothreitol. Methods Enzymol. *25*, 185-188.

Kozak,M. (1981). Possible role of flanking nucleotides in recognition of the AUG initiator codon by eukaryotic ribosomes. Nucl. Acids Res. *9*, 5233-5252.

Kozak,M. (1986). Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes. Cell *44*, 283-292.

Krishna,TG., Croy,RRD. and Boulter,D. (1979). Heterogeneity in subunit composition of the legumin of *Pisum sativum*. Phytochemistry *18*, 1879-1880.

Laemmli,UK. (1970). Cleavage of structural proteins during the assembly of the head of bacteriophage T4. Nature *227*, 680-685.

Lamport,DTA. (1969). The isolation and partial characterisation of hydroxyproline rich glycoproteins obtained by enzymic degredation of primary cell walls. Biochemistry *8*, 1155-1163.

Lamport,DTA. and Epstein,L. (1983). A new model for the primary cell wall: A concatenated extensin-cellulose network. Curr. Top. Plant Biochem. Physiol. *2*, 73-83.

Lamport,DTA., Katona,L. and Roerig,S. (1973). Galactosyl serine in extensin. Biochem. J. *133*, 125-131.

Lamprecht,H. (1941). On gene labillity in pea. Zuchter *13*, 97-107.

Lamprecht,H. (1953). New and hitherto known polymeric genes of *Pisum*. Agri. Hort. Genet. *11*, 40-54.

Lee,D., Ellis,THN., Turner,L., Hellens,RP. and Cleary,WG. (1990). A *copia*-like element in *Pisum* demonstrates the uses of dispersed repeated sequences. Plant Mol. Biol. *15*, 707-722.

Lee,M. and Macmillan,JD. (1970). Mode of action of pectic enzymes. III. Site of initial action of tomato pectinesterase on highly esterified pectin. Biochemistry *9*, 1930-1934.

Lelievre,J-M., Dickinson,CD., Dickinson,LA. and Nielsen,NC. (1992). Synthesis and assembly of soybean β-conglycinin *in vitro*. Plant Mol. Biol. *18*, 259-274.

Levasseur,MD. (1988). Comparative studies of the nucleotide sequences of pea seed storage protein genes. Ph.D. Thesis, Durham University.

Lindstrom,JT. and Vodkin,LO. (1991). A soybean cell wall protein is affected by seed colour genotype. Plant Cell *3*, 561-571.

Linn,F., Heidmann,I., Saedler,H. and Meyer,P. (1990). Epigenetic changes in the expression of the maize *A1* gene in *Petunia hybrida*: Role of numbers of integrated gene copies and state of methylation. Mol. Gen. Genet. *222*, 329-336.

Ludwig,SR. and Wessler,SR. (1990). Maize *R* gene family: Tissue-specific helix-loop-helix proteins. Cell *62*, 849-851.

Ludwig,SR., Habera,LF., Dellaporta,SL. and Wessler,SR. (1989). *Lc*, a member of the maize *R* gene family responsible for the tissue-specific anthocyanin production, encodes a protein similar to transcriptional activators and contains the *myc*-homology region. Proc. Natl. Acad. Sci. USA *86*, 7092-7096.

Lycett,GW., Delauney,AJ., Gatehouse,JA., Gilroy,J., Croy,RRD. and Boulter,D. (1983a). The vicilin gene family of pea (*Pisum sativum* L.): a complete cDNA coding sequence for preprovicilin. Nucl. Acids Res. *11*, 2367-2380.

Lycett,GW., Delauney,AJ. and Croy,RRD. (1983b). Are plant genes different? FEBS Lett. *153*, 43-46.

Lycett,GW., Croy,RRD., Shirsat,AH. and Boulter,D. (1984a). The complete nucleotide sequence of a legumin gene from pea (*Pisum sativum* L.). Nucl. Acids Res. *12*, 4493-4506.

Lycett,GW., Delauney,AJ., Zhao,W., Gatehouse,JA., Croy,RRD. and Boulter,D. (1984b). Two cDNA clones coding for the legumin protein of pea (*Pisum sativum* L.) contain sequence repeats. Plant Mol. Biol. *3*, 91-96.

Lycett,GW., Croy,RRD., Shirsat,AH., Richards,DM. and Boulter,D. (1985). The 5' flanking regions of three pea legumin genes; comparison of the DNA sequences. Nucl. Acids Res. *13*, 6733-6743.

Mahmoud,SH. (1985). Biochemical marker genes for molecular genetics and plant breeding in *Pisum sativum* L. Ph.D. Thesis, Durham University.

Maniatis,T., Fritsch,EF. and Sambrook,J. (1982). Molecular cloning - a laboratory manual. Cold Spring Harbor Laboratory, New York.

March,JF., Pappin,DJC. and Casey,R. (1988). Isolation and characterization of a minor legumin and its constituent polypeptides from *Pisum sativum* L. Biochem. J. *250*, 911-915.

Marcus,A., Greenberg,J. and Averyhart-Fullard,V. (1991). Repetitive proline-rich proteins in the extracellular matrix of the plant cell. Physiol. Plant. *81*, 273-279.

Markovic,O. and Jornvall,H. (1986). Pectinesterase: The primary structure of the tomato enzyme. Eur. J. Biochem. *158*, 455-462.

Martin,C., Prescott,A., Lister,C. and MacKay,S. (1989). Activity of the transposon *Tam3* in *Antirrhinum* and tobacco: possible role of DNA methylation. EMBO J. *8*, 997-1004.

Matta,NK. and Gatehouse,JA. (1982). Inheritance and mapping of storage protein genes in *Pisum sativum* L. Heredity *48*, 383-392.

Matta,NK., Gatehouse,JA. and Boulter,D. (1981). Molecular and subunit heterogeneity of legumin of *Pisum sativum* L. (garden pea) - A multi-dimesional gel electrophoretic study. J. Exp. Bot. *32*, 1295-1307.

McKnight,SL. and Kingsbury,RC. (1982). Transcriptional control signals of a eukaryotic protein coding gene. Science *217*, 316-324.

Meakin,PJ. and Roberts,JA. (1990). Dehiscence of fruit in oilseed rape (*Brassica napus* L.) II. The role of cell wall degrading enzymes and ethylene. J. Exp. Bot. *41*, 1003-1011.

Meakin,PJ. and Gatehouse,JA. (1991). Interaction of seed nuclear proteins with transcriptionally-enhancing regions of the pea (*Pisum sativum* L.) *legA* gene promoter. Planta *183*, 471-477.

Messing,J. (1983). New M13 vectors for cloning. Methods Enzymol. *101*, 20-78.

Messing,J., Geraghty,D., Heidecker,G., Hu,N., Kridl,J. and Robenstein, I. (1983). Plant gene structure. In, Genetic engineering of plants. (Kosuge,T., Meredith,CP. and Hollaender,A. eds.), pp211-227. Plenum, New York.

Mignery,GA., Pikaard,CS., Hannapel,DJ. and Park,WD. (1984). Isolation and sequence analysis of cDNAs for the major potato tuber protein, patatin. Nucl. Acids Res. *12*, 7987-8000.

Miller,K. (1987). Gel electrophoresis of RNA. Focus *9* (3), 14-15.

Morelli,G., Nagy,F., Fraley,RT., Rogers,SG. and Chua,N-H. (1985). A short conserved sequence is involved in the light-inducability of a gene encoding ribulose 1,5-bisphosphate carboxylase small subunit of pea. Nature *315*, 200-204.

Moustacas,A-M., Nari,J., Borel,M., Noat,G. and Ricard,J. (1991). Pectin methylesterase, metal ions and plant cell-wall extension. Biochem. J. *279*, 351-354.

Nasra,H. and Deacon,NJ. (1982). Relationship between the total size of exons and introns in protein-coding genes of higher eukaryotes. Proc. Natl. Acad. Sci. USA *79*, 6196-6200.

Nevers,P., Shepherd,NS. and Saedler,H. (1986). Plant transposable elements. Ad. Bot. Res. *12*, 103-203.

Newbigin,EJ., de Lumen,BO., Chandler,PM., Gould,A., Blagrove,RJ., March,JF., Kortt,AA. and Higgins,TJV. (1990). Pea convicilin: structure and primary sequence of the protein and expression of a gene in the seeds of transgenic tobacco. Planta *180*, 461-470.

Odani,S., Koide,T., Ono,T., Seto,Y. and Tanaka,T. (1987). Soybean hydrophobic protein. Isolation, partial characterization and the complete primary structure. Eur. J. Biochem. *162*, 485-491.

O'Reilly,C., Shepherd,NS., Pereira,A., Schwarz-Sommer,Z., Bertram,I., Robertson,DS., Peterson,PA. and Saedler,H. (1985). Molecular cloning of the *a1* locus of *zea mays* using the transposable elements *En* and *Mu1*. EMBO J. *4*, 877-882.

Osborne,TB. (1924). The vegetable proteins. Longmans, Green & Co., London.

Pang,S-Z., Rasmussen,J., Ye,G-N. and Sanford,JC. (1992). Use of the signal peptide of *Pisum* vicilin to translocate β-glucuronidase in *Nicotiana tabacum*. Gene *112*, 229-234.

Pearson,WR. and Lipman,DJ. (1988). Improved tools for biological sequence comparison. Proc. Natl. Acad. Sci. USA *85*, 2444-2448.

Pikaard,CS., Brusca,JS., Hannapel,DJ. and Park,WD. (1987). The two classes of genes for the major tuber protein patatin are differentially expressed in tubers and roots. Nucl. Acids Res. *15*, 1979-1994.

Plastow,GS. (1988). Molecular cloning and nucleotide sequence of the pectin methyl esterase gene of *Erwinia chrysanthemi* B374. Mol. Microbiol. *2*, 247-254.

Polhill,RM. (1981). Papilionoideae. In, Advances in legume systematics. (Polhill,RM. and Raven,PH. eds.), pp191-208. Royal Botanic Gardens, Kew, Richmond, Surrey, UK.

Pressey,R. and Reger,BJ. (1988). Polygalacturonase in pollen from corn and other grasses. Plant Sci. *59*, 57-62.

Ray,J., Knapp,J., Grierson,D., Bird,C. and Schuch,W. (1988). Identification and sequence determination of a cDNA clone for tomato pectin esterase. Eur. J. Biochem. *174*, 119-124.

Raz,R., Cretin,C., Puigomenech,P. and Martinez-Izquierdo,JA. (1991). The sequence of a hydroxyproline-rich glycoprotein from *Sorghum vulgare*. Plant Mol. Biol. *16*, 365-367.

Reddy,AR., Britsch,L., Salamini,F., Saedler,H. and Rhode,W. (1987). The *A1* (anthocyanin-1) locus in *Zea mays* encodes dihydroquercetin reductase. Plant Sci. *52*, 7-13.

Rees,DA. (1982). Polysacharide conformation in solutions and gels - recent results on pectins. Carbohydr. Polym. *2*, 254-263.

Rerie,WG., Wheitecross,MI. and Higgins,TJV. (1990). Nucleotide sequence of an A-type legumin gene from pea. Nucl. Acids Res. *18*, 655.

Rexova-Benkova,L. and Markovic,O. (1976). Pectic enzymes. Ad, Carbohydr. Chem. Biochem. *33*, 323-385.

Rhodes,PR. and Vodokin,LO. (1988). Organization of the *Tgm* family of transposable elements in soybean. Genetics *120*, 597-604.

Ricard,J. and Noat,G. (1986). Electrostatic effects and the dynamics of enzyme reactions at the surface of plant cells. I. A theory of the ionic control of a complex multienzyme system. Eur. J. Biochem. *155*, 183-190.

Richardson,M. (1991). Seed storage proteins: The enzyme inhibitors. In, Methods in plant biochemistry: Amino acids proteins and nucleic acids. (Rogers,LJ., Dey,PM. and Harborne,JB. eds.), pp259-305. Academic Press, New York.

Riggs,CD., Voelker,TA. and Chrispeels,MJ. (1989). Cotyledon nuclear proteins bind to DNA fragments harboring regulatory elements of phytohemagglutinin genes. Plant Cell *1*, 609-621.

Ryder,TB., Hedrick,SA., Bell,JN., Liang,X., Clouse,SD. and Lamb,CJ. (1987). Organization and differential activation of a gene family encoding the plant defense enzyme chalcone synthase in *Phaseolus vulgaris*. Mol. Gen. Genet. *210*, 219-233.

Salts,Y., Wachs,R., Gruissem,W. and Barg,R. (1991). Sequence coding for a novel proline rich protein preferentially expressed in young tomato fruit. Plant Mol. Biol. *17*, 149-150.

Salts,Y., Kenigsbuch,D., Wachs,R., Gruissem,W. and Barg,R. (1992). DNA sequence of tomato fruit expressed proline-rich gene TPRP-F1 reveals an intron within the 3' untranslated transcript. Plant Mol. Biol. *18*, 407-409.

Sanger,F., Nicklen,S. and Coulson,AR. (1977). DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. USA *74*, 5463-5467.

Sauer,N., Corbin,DR., Keller,B. and Lamb,CJ. (1990). Cloning and characterization of a wound-specific hydroxyproline-rich glycoprotein in *Phaseolus vulgaris*. Plant Cell Environ. *13*, 257-266.

Sawyer,RM. (1986). Isolation of a vicilin gene from pea (*Pisum sativum* L.), and nuclease sensitivity of seed storage protein genes in pea chromatin. Ph.D. Thesis, Durham University.

Scallon,BJ., Dickinson,CD. and Nielsen,NC. (1987). Characterization of a null-allele for the Gy_4 glycinin gene from soybean. Mol. Gen. Genet. *208*, 107-113.

Schwarz-Sommer,Z., Sheperd,N., Tacke,E., Gierl,A., Rohde,W. and Saedler,H. (1987). Influence of transposable elements on the structure and function of the *A1* gene of *Zea mays*. EMBO J. *6*, 287-294.

Shaw,CH., Carter,GH., Watson,MD. and Shaw,CH. (1984). A functional map of the nopaline synthase promoter. Nucl. Acids Res. *12*, 7831-7846.

Sheng,J., D'Ovidio,R. and Mehdy,MC. (1991). Negative and positive regulation of a novel proline-rich protein mRNA by fungal elicitor and wounding. Plant J. *1*, 345-354.

Shirsat,AH. (1988). A transposon-like structure in the 5' flanking sequence of a legumin gene from *Pisum sativum*. Mol. Gen. Genet. *212*, 129-133.

Shirsat,AH., Meakin,PJ. and Gatehouse,JA. (1990). Sequences 5' to the conserved 28bp Leg box element regulate the expression of pea seed storage protein *legA*. Plant Mol. Biol. *15*, 685-693.

Showalter,AM. and Varner,JE. (1988). Plant hydroxyproline-rich glycoproteins. in, The biochemistry of plants. (Stumpf,PK. and Conn,EE. eds.),vol 15, pp485-520. Academic Press, New York.

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

Spencer,D., Chandler,PM., Higgins,TJV., Inglis,AS. and Rubira,M. (1983). Sequence interrelationships of the subunits of vicilin from pea seeds. Plant Mol. Biol. *2*, 259-267.

Spok,A., Stubenrauch,G., Schorgendorfer,K. and Schwab,H. (1991). Molecular cloning and sequencing of a pectinesterase gene from *Pseudomonas solanacearum*. J. Gen. Microbiol. *137*, 131-140.

Staden,R. (1982). An interactive graphics program for computing and aligning nucleic acid and amino acid sequences. Nucl. Acids Res. *10*, 2951-2961.

Stafstrom,JP. and Staehelin,LA. (1988). Antibody localization of extensin in cell walls of carrot storage roots. Planta *174*, 321-332.

Statham,CM. and Murfet,IC. (1974). Inheritance of flavone production in *Pisum* flowers. Pisum News Lett. *6*, 49.

Statham,CM., Crowden,RK. and Harborne,JB. (1972). Biochemical genetics of pigmentation in *Pisum sativum*. Phytochemistry *11*, 1083-1088.

Stiefel,V., Perez-Grau,L., Albertico,F., Giralt,E., Ruiz Avila,L., Ludevid,MD. and Puigdomenech,P. (1988). Molecular cloning of a cDNA encoding a putative cell wall protein from *Zea mays* and immunological identification of related polypeptides. Plant Mol. Biol. *11*, 483-493.

Stiefel,V., Ruiz-Avila,L., Raz,R., Valles,MP., Gomez,J., Pages,M., Martinez-Izquierdo,JA., Ludevid,MD., Langdale,JA., Nelson,T. and Puigdomenech,P. (1990). Expression of a maize cell wall hydroxyproline-rich glycoprotein gene in early leaf and root vascular differentiation. Plant Cell *2*, 785-793.

Summers,WL. (1989). Pectinesterase and D-galacturonase activities in eight snap bean cultivars. Hortscience *24*, 484-486.

Talbott,LD. and Ray,PM. (1992). Molecular size and seperability features of pea cell wall polysaccharides: Implications for models of primary wall structure. Plant Physiol. *98*, 357-368.

Thomas,PS. and Farquhar,MN. (1978). Specific measurement of DNA in nuclei and nucleic acids using Diaminobenzoic acid. Anal. Biochem. *89*, 35-44.

Thompson,AJ., Evans,IM., Boulter,D., Croy,RRD. and Gatehouse,JA. (1989). Transcriptional and posttranscriptional regulation of seed storage-protein gene expression in pea (*Pisum sativum* L.). Planta *179*, 279-287.

Thompson,AJ., Bown,D., Yaish,S. and Gatehouse,JA. (1991). Differential expression of seed storage protein genes in the pea *legJ* subfamily; sequence of gene *legK*. Biochem. Pysiol. Pflanzen *187*, 1-12.

Thomson,JA. and Schroeder,HE. (1978). Colyledonary storage proteins in *Pisum sativum*. II. Hereditary variation in components of the legumin and vicilin fractions. Aust. J. Plant Physiol. *5*, 281-294.

Thomson,JA., Schroeder,HE. and Tassie,AM. (1980). Cotyledonary storage proteins in *Pisum sativum*. V. Further studies in molecular heterogeneity in the vicilin series of holoproteins. Aust. J. Plant Physiol. *7*, 271-282.

Tierney,ML., Wiechert,J. and Pluymers,D. (1988). Analysis of the expression of extensin and p33-related cell wall proteins in carrot and soybean. Mol. Gen. Genet. *211*, 393-399.

Varner,JE. and Hood,EE. (1988). Gel properties of the cell wall. In, Self-Assembling architecture. (Varner,JE. ed.), pp97-103. AR.Liss, New York.

Varner,JE. and Lin,L-S. (1989). Pant cell wall architecture. Cell *56*, 231-239.

Vodkin,LO. (1989). Transposable element influence on plant gene expression and variation. In, Biochemistry of plants. (Stumpf,PK. and Conn,EE. eds.), vol.15, pp83-123. Academic Press, London.

Vodkin,LO., Rhodes,PR. and Goldberg,RB. (1983). A lectin gene insertion has the structural features of a transposable element. Cell *34*, 1023-1031.

Voelker,TA., Moreno,J. and Chrispeels,MJ. (1990). Expression analysis of a pseudogene in transgenic tobacco: a frameshift mutation prevents mRNA accumulation. Plant Cell *2*, 255-261.

Von Heijne,G. (1983). Patterns of amino acids near signal-sequence cleavage sites. Eur. J. Biochem. *133*, 17-21.

Weschke,W., Baumlein,H. and Wobus,U. (1987). Nucleotide sequence of a field bean (*Vicia faba* L. var. minor) vicilin gene. Nucl. Acids Res. *15*, 10065.

Wessler,SR. (1988). Phenotypic diversity mediated by the maize transposable elements *Ac* and *Spm*. Science *242*, 399-405.

Wiel,CF. and Wessler,SR. (1990). The effects of plant transposable element insertion on transcription initiation and RNA processing. Annu. Rev. Plant Physiol., Plant Mol. Biol. *41*, 527-552.

Wienand,U., Wegdemann,U., Niesbach-Klosgen,U., Peterson,PA. and Saedler,H (1986). Molecular cloning of the *c2* locus of *Zea mays*, the gene coding for chalcone synthase. Mol. Gen. Genet. *203*, 202-207.

Wing,RA., Yamaguchi,J. and Larabell,SK. (1990). Molecular and genetical characterisation of two pollen-expressed genes that have sequence similarity to pectate lyases of the plant pathogen *Erwinia*. Plant Mol. Biol. *14*, 17-28.

Wingender,R., Rohrig,H., Horicke,C., Wing,D. and Schell,J. (1989). Differential regulation of soybean chalcone synthase genes in plant defence, symbiosis and upon environmental stimuli. Mol. Gen. Genet. *218*, 315-322.

Woods,D (1984). Oligonucleotide screening of cDNA libraries. Focus *6* (3), 1-3.

Wyatt,RE., Nagao,RT. and Key,JL. (1992). Patterns of soybean proline-rich protein gene expression. Plant Cell *4*, 99-110.

Yaish,SA. (1990). Construction and screening of plant genomic libraries. Ph.D. Thesis, Durham University.

Yamaoka,T. and Chiba,N. (1983). Changes in the coagulation ability of pectin during growth of soybean hypocotyls. Plant Cell Physiol. *24*, 1281-1290.

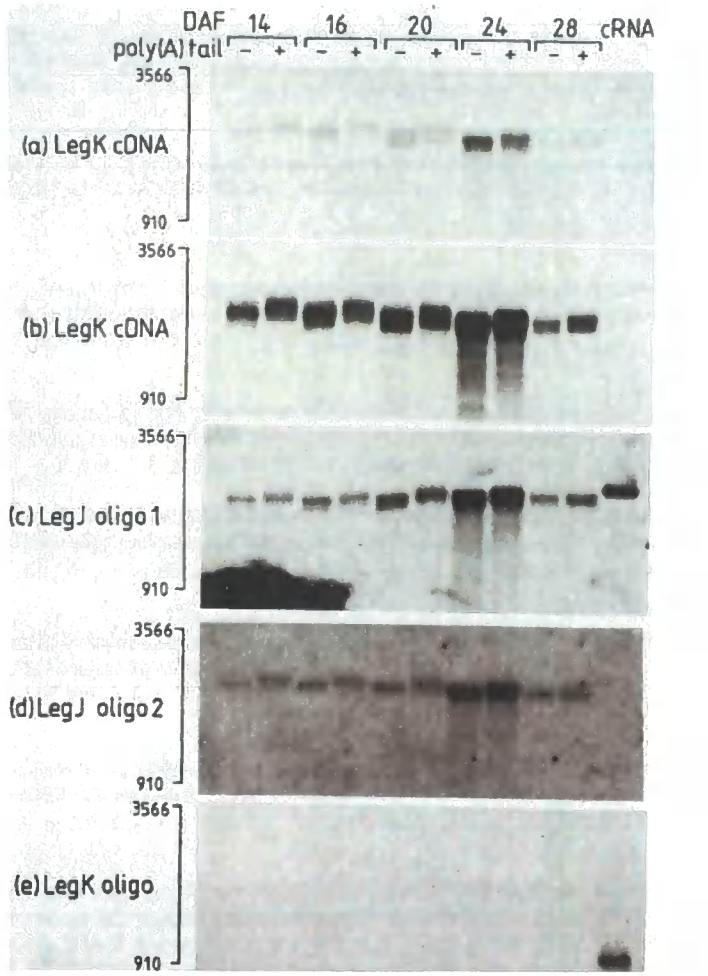
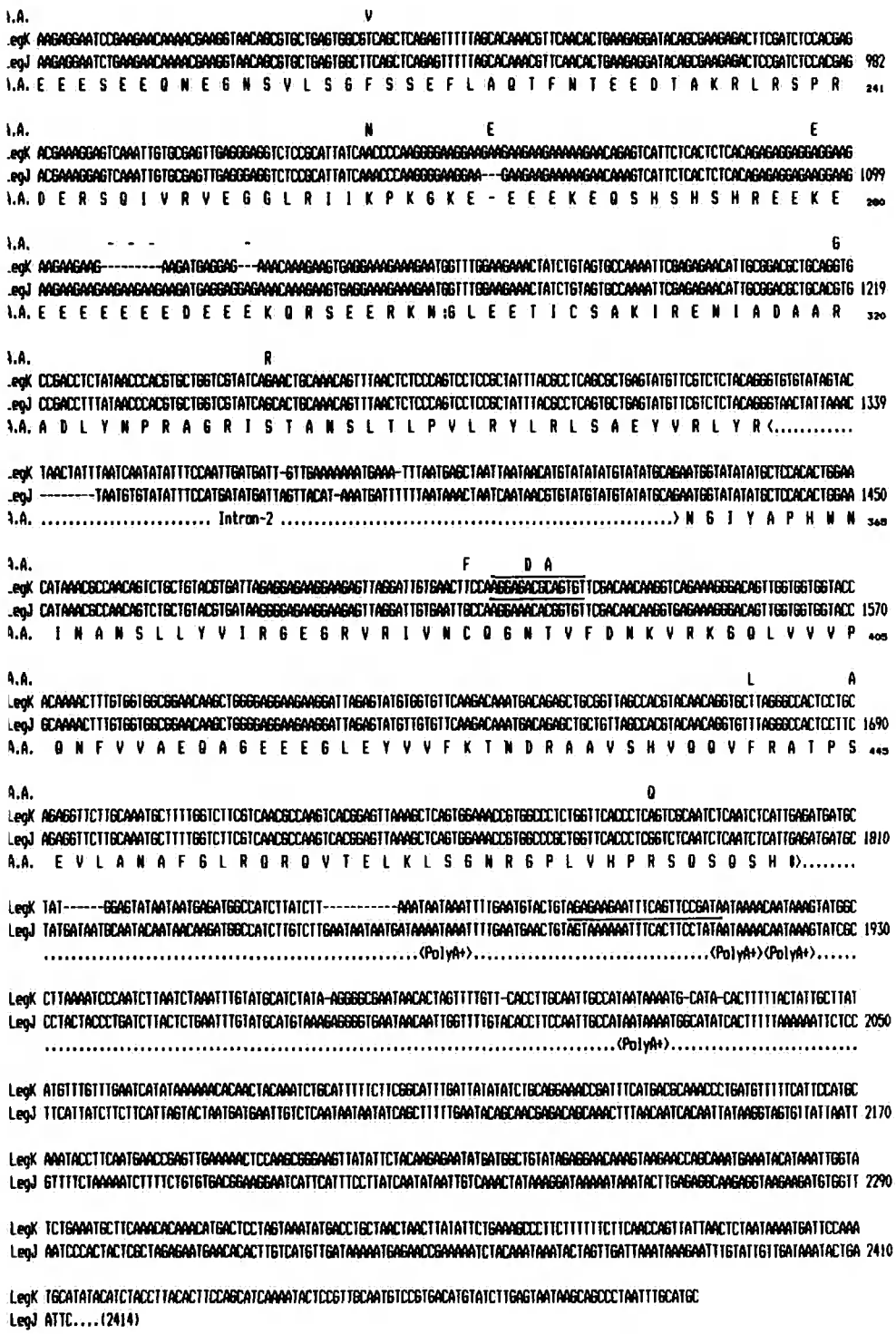
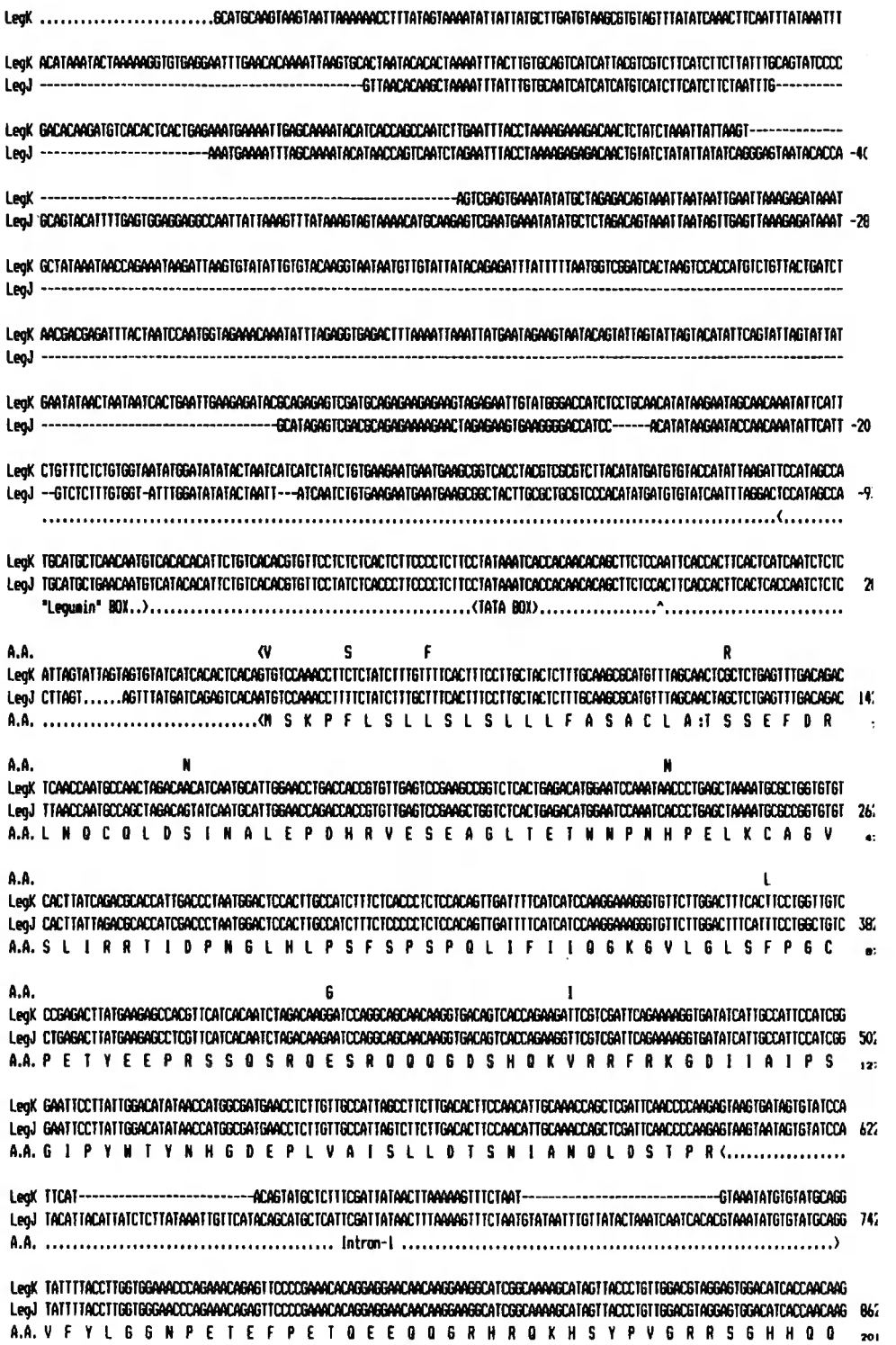


Fig. 3. A. *Analysis of the expression of the legJ subfamily by Northern blotting.* Total cotyledon RNA from the indicated days after flowering (DAF) were treated with oligo(dT)$_{12-18}$ and RNaseH, to remove the poly(A) tail. 1 µg samples of RNA, with and without poly(A) tail, labelled – and + respectively, were run on agarose-formaldehyde gels and Northern blotted. The blots were probed as follows: (a), (b), *legK* cDNA (pCD40), 5 h and 16 h exposure respectively (detects expression of all genes in the *legJ* subfamily); (c), *legJ* oligo1; (d) *legJ* oligo2; (e) *legK* oligo. Blots (c) and (e) also included, in tracks labelled cRNA, 1 fmol each of synthetic RNAs produced by transcribing inserts from *legJ* and *legK* coding and 3′ flanking sequence (plasmids pJC5.2 and pCD40 respectively), cloned into the expression vector pGEM-blue (Promega Biotech), using T7 RNA polymerase *in vitro*. Vertical axes represent size markers (pea rRNA and pBR322 Alu I DNA restriction fragments).

sequence oligo probes show that both bands of higher mobility are products of *legJ*, and that no mRNA corresponding to *legK* can be detected.

When the 3′ flanking sequence oligo, *legJ* oligo2, was used as probe, only 1 of the 2 bands detected by the *legJ* coding sequence oligo1 hybridised. This was the band at 72 mm (Fig. 3A (d) and Fig. 3B (c)).





rticular, the "legumin box" (BÄUMLEIN et al. 1986; GATEHOUSE et al. 1986) shows only 1 se difference between the 2 genes.

Expression of Members of the legJ Gene Subfamily in Developing Pea Cotyledons

Hybridisation of coding sequences of *legJ* and *legK*, and cDNA species derived from ɜm, to total RNA from developing pea cotyledons has previously been shown to detect 2 -defined bands of size approx. 1660 and 1860 nucleotides (GATEHOUSE et al. 1988). In der to improve the resolution of the bands, RNA samples were hybridised to oligo(dT)$_{12-18}$ d then incubated with RNaseH to remove poly(A) tails which may contribute to size terogeneity. Analysis of total RNAs with and without this treatment by Northern blotting is own in Fig. 3A. Probing such a blot with the complete *legK* coding sequence, as shown in g. 3A (a) and (b) shows that the treatment with RNaseH reduced the mean size of the 'bridising region by approx. 100 nucleotides, consistent with the removal of poly(A) tails. ıe RNaseH-treated samples show 3 clear hybridising bands, with a fourth present at low tensity; this is more clearly visualised on the densitometric scans of the blot shown in g. 3B (a). The densitometric scans clearly show that the different bands fall into 2 groups; bands of lower mobility (approx. 68 mm and 70 mm in Fig. 3B (a)), maintained at an ıproximately constant ratio, which are present at maximum amount at 16 daf (i. e. towards e end of the cotyledon expansion phase), and 2 bands of higher mobility (approx. 72 mm ıd 74 mm), again maintained at an approximately constant ratio, which are present at aximum amount at 24 daf (i.e. during the desiccation phase). The lower mobility bands us represent mRNA species which accumulate and decay approximately 8 d earlier in ıtyledon development than the higher mobility bands.

Identification of mRNA Species

The *legK* cDNA probe used to obtain the data described above has previously been ıown to hybridise to all 3 members of the *legJ* gene subfamily with approximately equal ficiency at high stringency, and thus the above results represent the total expression of ɜnes in this subfamily (DOMONEY and CASEY 1985). In order to distinguish different embers of the gene family more specific probes are required. It had previously been shown at the 3′ non-coding region of *legJ* hybridised to the more mobile band in untreated RNA ɜATEHOUSE et al. 1988). Oligonucleotides were synthesised complementary to regions of ɜquence that diverged between genes *legJ* and *legK* in order to distinguish mRNA species oduced by the different genes; these oligos covered bases 1516–1530 in *legJ* and the ɔrresponding region in *legK* (coding sequence; *legJ* oligol, 5′-ACACcGtGTtTCCT-3′, and *gK* oligo, 5′-ACACtGcGTcTCCT-3′; small letters indicate mismatched bases between the genes), and bases 1988–1910 in *legJ* (3′ flanking sequence; *legJ* oligo), '-TATaGGAAgTGAAATTtTTtaCT-3′; small letters indicate mismatches with correspondıg *legK* sequence). The specificity of these oligos was checked by hybridisation to synthetic NA species. Hybridisation of *legJ* oligo1 is shown in Fig. 3A (c) and 3B (b); binding only kes place to the higher mobility bands, i.e. those at 72 and 74 mm. No hybridisation above ackground was observed with the *legK* oligo, as shown in Fig. 3A (e). Thus the coding

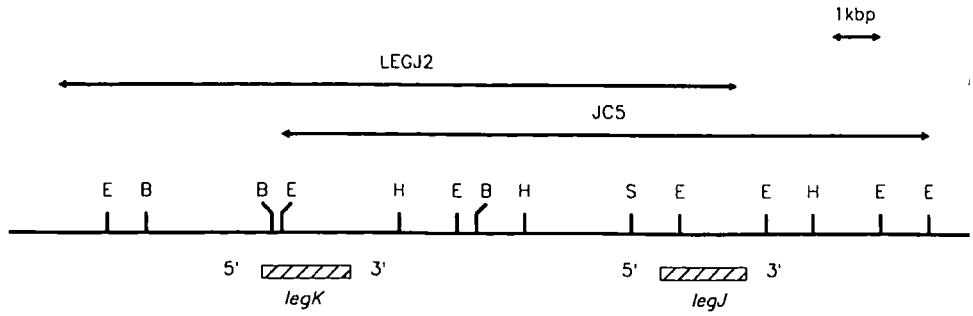


Fig. 1. *Partial restriction map of the region of pea genomic DNA covered by genomic clone lambda JC5 (Gatehouse et al. 1988) and LEGJ2.* Positions of the coding sequences of genes *leg* and *legK* are indicated. B = *Bam HI;* E = *Eco RI;* H = *Hind III;* S = *Sal I.*

entire sequence of gene *legK*, which had previously been partially characterised via th 3′ portion of its sequence present on genomic clone JC5; the 2 genomic clones overla to cover a region of approx. 17 kb.

The determined sequence of *legK* and its 3′ and 5′ flanking regions is given in Fig 2. Despite the 2 genomic clones having been isolated from different pea lines (JC5 wa isolated from pea line Dark Skinned Perfection) no differences between the previousl reported partial sequence of *legK* (from Dark Skinned Perfection) and the new complet *legK* sequence (from Feltham First) were observed.

The sequence of *legK* shows a high degree of homology to *legJ*, with only 2(amino acid substitutions, 1 addition and 4 deletions over a total of 482 amino acids However, one substitution in *legK* is of primary significance; the start codon ATG i *legJ* is mutated to GTG, a valine codon. The first subsequent ATG in *legK* is out o frame relative to *legJ*, at base 117 (*legJ* numbering), and gives an open reading fram of only 6 amino acids. The other changes in the *legK* coding sequence are within th normal range of variation observed in storage proteins of this type. Significant homol ogy between the 3′ flanking sequences of *legJ* and *legK* extends for approx. 250 base after the stop codons, with a sharp "cut-off" approx. 20 bases 3′ to the fourth poly adenylation signal sequence; in contrast, homology between the 5′ flanking sequence of the 2 genes is significant over the entire region determined (to approx. −560 base in legJ), although each gene has sizeable insertions relative to the other (Fig. 2). Con served sequence elements in the *legJ* 5′ flanking sequence are also present in *legK;* i

Fig. 2. *Sequences of genes legJ (Gatehouse et al. 1988) and legK, aligned for maximum homol ogy.* Base 1 in *legJ* is indicated by a circumflex (^). Polyadenylation signal sequences are indi cated by ‹PolyA+›. Sequences complementary to oligonucleotides *legJ* oligos 1 and 2, and *legJ* oligo are overlined. Other features are as indicated. Amino acids are only given for *legK* wher they differ from *legJ*. The 5′ flanking sequences of the genes, and the 3′ flanking sequences t base 2039 in *legJ*, are aligned for maximum homology. There is no significant homology in th 3′ flanking sequences of the genes beyond base 2039 in *legJ*.

abelled deoxy- and dideoxy-nucleotides; *Taq* polymerase was used in the extension reactions, ording to the protocols supplied by Applied Biosystems. Sequences were determined on an Applied systems Model 370A automated DNA sequencer. Both strands of the DNA were fully sequenced. juence analysis was carried out using software kindly supplied by Dr. W. PEARSON.

gonucleotides

They were produced on an Applied Biosystems Model 381A DNA synthesiser and purified by erse phase chromatography. 10 pmol amounts were end-labelled using 100 μCi (20 pmol) of [γ-]ATP under standard reaction conditions (40 mM Tris-HCl, *p*H 7.5, 10 mM $MgCl_2$, 5 mM DTT, 5 ts T4 polynucleotide kinase; incubate at 37°C for 60 min), and unincorporated nucleotide was ıoved by gel filtration on a column of Sephadex G-25.

raction and analysis of RNA

Total RNA was extracted from pea tissues by the "hot SDS" method of HALL et al. (1978). Tissues :e frozen in liquid air immediately on harvesting and stored at −80°C. RNA was analysed by :trophoresis through denaturing agarose gels containing 2.2 M formaldehyde (MILLER 1987); DNA ! RNA size markers were used interchangeably).

noval of poly(A) sequences from RNA

Poly(A) sequences were removed from mRNA by treatment with RNaseH in the presence of ;o$(dT)_{12-18}$, as described by VOURNAKIS et al. (1975). 12 μg of total RNA was incubated in 25 mM Cl_2, 20 mM Tris-HCl *p*H 8.0 with 2 μg of oligo$(dT)_{12-18}$ at 25°C for 15 min in a volume of 40 μl. units of RNaseH were added, and the solution was incubated at 37°C for 20 min. RNA was ɔvered by 2 phenol/chloroform extractions and ethanol precipitation.

'thern blotting

After electrophoresis, RNA was blotted onto nylon membranes (Hybond-N, Amersham Interna-ıal plc) using 20×SSC (1×SSC is 0.15 M NaCl, 0.015 M sodium citrate, *p*H 7.2) as transfer fer. RNA was fixed onto the membranes after blotting by uv-crosslinking (2 min exposure to 300 uv on a transilluminator), followed by baking *in vacuo* at 80°C for 20 min. Blots were hybridised in 20 ml of 50% deionised formamide containing 0.5 M NaCl, 1 mM EDTA, 40 mM 'ES-NaOH *p*H 6.5, 0.4% SDS, 100 μg · ml^{-1} denatured herring sperm DNA, 100 μg · ml^{-1} yadenylic acid and 5×Denhardt's solution, at 41°C for 8−16 h. DNA probes were produced by dom-primed labelling of isolated DNA fragments with [α-^{32}P]dCTP, separated from unincorporated :leotide by gel filtration on Sephadex G-50, denatured, and added to the blot in a volume of 0.8−1.2 After hybridisation for 40−48 h blots were washed at 60°C in 1.4×SSC, 0.1% SDS for 4×30 min shes, then 2×30 min washes in 0.1×SSC. Blots were then autoradiographed while wet. gonucleotide probes were hybridised to blots under similar conditions, except that the solution ıtained 0.9 M NaCl, 90 mM Tris-HCl *p*H 7.5 as buffer and no Denhardt's solution, at a lower ıperature (T_m-8°C) predicted by the T_m for the oligonucleotide. Washing was to a stringency of T_m-: in 6×SSC, 0.1% SDS. Blots were autoradiographed at −80°C using flashed x-ray film (Fuji RX) l an intensifying screen (DuPont Lightning Plus). Autoradiographs were analysed visually and by ısitometric scanning on an LKB Ultroscan XL densitometer.

Results

Isolation, Characterisation and Sequence of legK

A partial restriction map of the region of pea genomic DNA covered by the clones ıbda JC5 and LEGJ2 is given in Fig. 1. The newly isolated clone, LEGJ2, contains the

evidence that *legA* must be expressed as it is active in transgenic plants. The very high degrees of homology of the sequences of genes in this subfamily (> 99% in coding sequence) has made measurements of individual gene expression by nucleic acid hybridisation difficult. These genes, and the polypeptides they encode, have been mapped to a single genetic locus, *Lg-1*, near to *r* on chromosome 7 of the pea genome (MATTA and GATEHOUSE 1982; DOMONEY et al. 1986b). In contrast, the "minor" subunits did not segregate as a single locus (CASEY 1979), and subsequent investigation showed the presence of 3 distinct subfamilies of genes encoding these polypeptides (THOMPSON 1989). Of these "minor" legumin gene subfamilies, one is fairly well characterised. It contains 3 genes; a fully characterised and sequentced gene, *legJ*, and part of a second gene, *legK*, were found on the same genomic clone (GATEHOUSE et al. 1988). The third gene, tentatively designated *legL*, has been identified by Southern blots of pea genomic DNA whith subfamily-specific probes (DOMONEY and CASEY 1985). The *legJ* subfamily genes map to a locus, *Lg-2*, near *a* on chromosome 1 of the pea genome (DOMONEY et al. 1986b). The differences in sequence between the genes of the *legJ* subfamily are greater than between those in the *legA* subfamily; the present paper reports experiments that distinguish the expression patterns of single genes within the *legJ* subfamily.

Materials and Methods

Plant material

Pea (*Pisum sativum* L., cv. Feltham First; Suttons Seeds, Torquay, U.K.) plants were grown hydroponically in pots containing Phostrogen (0.55 g · l^{-1}) under standardised conditions as previously described (EVANS et al. 1979; THOMPSON 1989).

Reagents and enzymes

Unless otherwise specified, DNA restriction and modification enzymes were from Northumbrian Biologicals Ltd. (Cramlington, U.K). DNA sequencing reagents were from Applied Biosystems (Warrington, U.K.).
Radiochemicals were supplied by Amersham International plc, as was nylon membrane for Northern blotting. Other reagents were of analytical quality, or best grade available.

Isolation and characterisation of genomic clone

The genomic clone LEGJ2 was isolated from a genomic library of Sau3A I fragments of pea (cv. Feltham First) DNA, produced by partial restriction and size-selection, cloned in the lambda vector EMBL3; this library was the gift of Dr. A. H. SHIRSAT, and has been described previously (LYCETT et al. 1986). The library was screened with a probe (pJC5.2; GATEHOUSE et al. 1988) corresponding to the coding sequence of gene *legJ;* positive plaques on duplicate screens were plaque-purified. The clone giving the strongest hybridisation signal was selected for further study, and was designated LEGJ2. DNA was prepared from this phage, and was restriction-mapped. Selected fragments were subcloned into pUC18 and 19 plasmid vectors for further analysis.

DNA sequencing

DNA fragments produced by restriction enzyme digestion were subcloned in M13mp18 and 19 vectors, and used to produce single stranded templates for sequencing. DNA sequencing was carried out by the dideoxynucleotide chain termination method, using fluorescent labelled primers and

iochem. Physiol. Pflanzen **187**, 1–12 (1991)
ustav Fischer Verlag Jena

)ifferential Expression of Seed Storage Protein Genes in the Pea *legJ* Subfamily; Sequence of Gene *legK*

ANDREW, J. THOMPSON, DAVID BOWN, SAMI YAISH and JOHN A. GATEHOUSE

Department of Biological Sciences, University of Durham, Durham, Great Britian

.ey Term Index: DNA sequence, gene regulation, seed storage proteins, *Pisum sativum* L.

Summary

The *legJ* subfamily of genes in garden pea, *Pisum sativum* L., encodes "minor" legumin seed :orage protein polypeptides. Data on the differential expression of the 3 genes *(legJ,K,L)* within this ıbfamily is reported. The expression of one gene *(legJ)* is specifically upregulated during the esiccation phase of cotyledon development, when other storage protein genes are downregulated. The ɔmplete sequence of a second gene in the subfamily, *legK*, shows that the failure to observe any ‹pression of this gene is due to the mutation of its initiator ATG (methionine) codon to a GTG (valine) ɔdon. The third gene in the subfamily, *legL*, shows maximal expression during the cotyledon ‹pansion phase of seed development, i.e. like other storage protein genes. Evidence for the use of ltemative polyadenylation addition signal sequences in these genes is also presented.

Introduction

Legumin is a major seed protein in pea (*Pisum sativum* L.), and is representative of a type f storage protein widely distributed throughout the plant kingdom. The protein is the roduct of a multi-gene family, and is inherently heterogeneous, like many other storage roteins. The basic unit of legumin is a disulphide-bonded subunit pair, synthesised as a ingle precursor polypeptide of approx. 500 amino acids. The precursor contains, in order rom its N-terminus, a leader sequence which directs its transport into the endoplasmic ɜticulum, from where the protein is transported to the storage organelles (protein bodies) via ıe Golgi apparatus, the α- or acidic polypeptide (usually approx. 40,000 M_r) and the β- or ›asic polypeptide (usually approx. 20,000 M_r); the formation of the disulphide bond between ıe α- and β-polypeptides precedes their separation, so that specific α-polypeptides are lways linked to specific β-polypeptides. The leader sequence is removed in a cotranslational ›roteolytic cleavage, and the α- and β-polypeptides are separated by a post translational ›roteolytic cleavage that probably takes place in the protein bodies. Subunit pairs in pea are ssembled into hexameric molecules, M_r 380–400,000, which can be homo- or heteropoly-ıers (reviewed by CROY and GATEHOUSE 1985).

Legumin subunit pairs have been classified into 2 types, "major" and "minor" on the ›asis of the relative abundances of their α-polypeptides (CASEY 1979). The "major" subunit ıairs were subsequently shown to be the products of a highly homologous sub-family of ;enes designated *legA-E*, which have been characterised and sequenced (LYCETT et al. 1984; ,YCETT et al. 1986; SHIRSAT, A. H. and GATEHOUSE, J. A., unpublished results). No data ›n expression of individual genes in the *legA* subfamily has been obtained, apart from

EVANS, I. M., CROY, R. R. D., HUTCHINSON, P., BOULTER, D., PAYNE, P. I., and GORDON, M. E.: Cell free synthesis of some storage protein subunits by polyribosomes and RNA from developing seeds of pea (*Pisum sativum* L.). Planta **144**, 455–462 (1979).

GATEHOUSE, J. A., EVANS, I. M., CROY, R. R. D., and BOULTER, D.: Differential expression of genes during legume seed development. Phil. Trans. Roy. Soc. Lond. **B314**, 367–384 (1986).

GATEHOUSE, J. A., BOWN, D., GILROY, J., LEVASSEUR, M., CASTLETON, J., and ELLIS, T. H. N.: Two genes encoding "minor" legumin polypeptides in pea (*Pisum sativum* L.). Biochem. J. **250**, 15–24 (1988).

HALL, T. C., MA, Y., BUCHBINDER, B. U., PAYNE, J. W., SUN, S. M. and BLISS, F. A.: Messenger RNA for G1 protein of french bean seeds: cell-free translation and product characterisation. Proc. Natl. Acad. Sci. USA **75**, 3196–3200 (1978).

LYCETT, G. W., CROY, R. R. D., SHIRSAT, A. H., and BOULTER, D.: The complete nucleotide sequence of a legumin gene from pea (*Pisum sativum* L.). Nucl. Acids Res. **12**, 4493–4506 (1984).

LYCETT, G. W., CROY, R. R. D., SHIRSAT, A. H., RICHARDS, D. M., and BOULTER, D.: The 5′-flanking regions of three pea legumin genes; comparison of the DNA sequences. Nucl. Acids Res. **13**, 6733–6743 (1986).

MATTA, N. K., and GATEHOUSE, J. A.: Inheritance and mapping of storage protein genes in *Pisum sativum* L. Heredity **48**, 383–392 (1982).

MILLER, K.: Gel electrophoresis of RNA. Focus (Bethesda Res. Labs) **9: 3**, 14–15 (1987).

SCALLON, B. J., DICKINSON, C. D., and NIELSEN, N. C.: Characterisation of a null-allele for the Gy_4 glycinin gene from soybean. Mol. Gen. Genet. **208**, 107–113 (1987).

THOMPSON, A. J.: Regulation of gene expression in developing pea seeds. Ph. D. thesis, University of Durham 1989.

THOMPSON, A. J., EVANS, I. M., BOULTER, D., CROY, R. R. D., and GATEHOUSE, J. A.: Transcriptional and posttranscriptional regulation of seed storage protein gene expression in pea (*Pisum sativum* L.). Plants **179**, 279–287 (1989).

VOELKER, T. A., MORENO, J., and CHRISPEELS, M. J.: Expression analysis of a pseudogene in transgenic tobacco: a frameshift mutation prevents mRNA accumulation. Plant Cell **2**, 255–261 (1990).

VOURNAKIS, J. N., EFSTRATIADIS, A., and KAFATOS, F. C.: Electrophoretic patterns of deadenylated chorion and globin mRNAs. Proc. Natl. Acad. Sci. USA **72**, 2959–2963 (1975).

Received May 14, 1990; *accepted* July 27, 1990.

Authors' addresses: Dr. J. A. GATEHOUSE, Mr. D. BOWN, Mr. S. YAISH, Department of Biological Sciences, University of Durham, South Road, Durham DH1 3LE, U.K. Dr. A. J. THOMPSON; Department of Botany. The University of Texas, Austin, Texas 78713-7640, U.S.A.

The observation that the single *legJ* gene produces multiple RNA species on Northern blots, even after removal of the heterogeneity caused by different poly(A) tail lengths, can be shown to be due to the use of different polyadenylation signal sequences. *LegJ* oligo 2 will only detect mRNA species that continue beyond base 1910; i.e. it will detect mRNA species that use the polyadenylation signals at bases 1909–1923, and 2015–2020, but not those that use the polyadenylation signal at bases 1869–1874 (which terminate approx. 20 bases after the poly(A) signal sequence). In contrast, *LegJ* oligo1 will detect all *legJ* mRNA species. The data presented show that *legJ* oligo2 detects only the lower mobility band (72 mm) of the 2 bands detected by *legJ* oligo1 (72 and 74 mm); thus the higher mobility (74 mm) band, corresponding to a shorter RNA, must represent mRNA species that use the first polyadenylation signal sequence, at 1869–1874. The observed difference in mobility suggests that the 72 mm band represents RNA species using the last poly(A) signal sequence, at 2015–2020. There is not a strong prefence for one or the other signal sequence, although the 3′ one is slightly favoured, nor does the preference change significantly with stage of seed development, and is thus unlikely to represent a mechanism for controlling expression. The occurrence of multiple poly(A) signal sequences in plant genes is common (DEAN et al. 1986), and use of alternative poly(A) signal sequences may not be uncommon; a minor legumin gene in a different subfamily, designated *legS*, has been shown to use alternative poly(A) signal sequences by sequence data form different cDNA clones (DOMONEY et al. 1986a; GATEHOUSE, J. A., unpublished data). Additionally, the multiple RNA species observed in Northern blots as products of *legL* in the present paper suggest that it also uses alternative signal sequences.

Acknowledgements

We thank Dr. ANIL SHIRSAT for making available a pea genomic library, Dr. MARTA EVANS for advice, and Miss JULIA BRYDEN for technical assistance with DNA sequencing.

Financial support from SERC (AJT), AFRC (JAG) and the British Council (SY) is gratefully acknowledged. We also thank Prof. D. BOULTER for departmental facilities.

References

BÄUMLEIN, H., WOBUS, U., PUSTELL, J., and KAFATOS, F. C.: The legumin gene family of *Vicia faba:* a B-type gene and a possible legumin gene regulatory element. Nucl. Acids Res. **14**, 2707–2720 (1986).

CASEY, R.: Genetic variability in the structure of the alpha-subunits of legumin from *Pisum* – a two dimensional gel electrophoretic study. Heredity **43**, 265–272 (1979).

CROY, R. R. D., and GATEHOUSE, J. A.: Genetic engineering of seed proteins: current and potential applications. In: "Plant Genetic Engineering" (Ed. DODDS, J. H.) Cambridge University Press, Cambridge 1985.

DEAN, C., TAMAKI, S., DUNSMUIR, P., FAVREAU, M., KATAYAMA, C., DOONER, H., and BEDBROOK, J.: mRNA transcripts of several plant genes are polyadenylated at multiple sites *in vivo*. Nucl. Acids Res. **14**, 2229–2240 (1986).

DOMONEY, C., and CASEY, R.: Measurement of gene number for seed storage proteins in *Pisum*. Nucl. Acids Res. **13**, 687–699 (1985).

DOMONEY, C., BARKER, R., and CASEY, R.: The complete deduced amino acid sequences of legumin beta-polypeptides from different genetic loci in *Pisum*. Plant Mol. Biol. **7**, 467–474 (1986a).

DOMONEY, C., ELLIS, T. H. N., and DAVIES, D. R.: Organisation and mapping of legumin genes in *Pisum*. Mol. Gen. Genetics **202**, 280–285 (1986b).

member of this gene subfamily whose expression is increased during the desiccation phase. The increase in amount *legJ* mRNA as a proportion of total RNA can be estimated as approx. 8-fold between 14 daf (midexpansion phase) and 24 daf (desiccation phase); as desiccation proceeds, expression of *legJ* decreases again, so that the proportion of mRNA has fallen by approx. 4-fold from 24 to 28 daf.

Discussion

The failure to observe expression from gene *legK* is explained by the mutation of its start codon to a GTG valine codon. Under these circumstances, an mRNA produced by transcription of *legK* will be translationally ineffective, and will consequently have a very short half-life, and a low steady state level. Recent experiments with the lectin gene *Pdlec1* from *Phaseolus vulgaris* (VOELKER et al. 1990) in transgenic plants have shown that the presence of stop condons in the reading frame of an mRNA decrease its steady state level 40-fold; the failure to detect *legK* mRNA under the conditions used in the present paper suggests a decrease in its steady state level, relative to *legJ*, at least as great as this. The presence of a small insertion in the *legK* sequence (relative to *legJ*) in the region where the ribosome would be expected to bind (base +34) may also affect the stability of the *legK* mRNA. Evidence from experiments with other similarly "damaged" genes suggests that the promotor sequence of *legK* should be active, so that the gene is expressed, but no gene product accumulates; the maintenance of the homology between the 5′ flanking sequences of *legJ* and *legK* supports this assumption. Further, *legK* must be expressed in other pea lines, since the *legK* cDNA pCD40 must represent a viable mRNA; this cDNA was produced from pea line "Birte". The mutation in *legK*, preventing its gene product from accumulating, present in pea line Feltham First but presumably absent in Birte, can account for some of the line-line variation seen in minor legumin α-polypeptides. Interestingly, an analogous mutation has been observed in the legumin (glycinin) genes of soya bean; a null allele of the Gy_4 gene observed at the protein level was correlated with a change in the start codon of the mRNA from ATG to ATA (SCALLON et al. 1987).

Since no mRNA from gene *legK* can be detected in pea line Feltham First the observed mRNA species must be produced by 2 genes only, *legJ* and the as yet uncharacterised *legL*. No other genes of sufficient homology are present on the pea genome to give cross-hybridising mRNA species. Since the *legJ* oligos have clearly identified the higher mobility RNA species as products of *legJ*, the lower mobility RNA species, at 68 and 70 mm, must be products of *legL*. Interestingly, whereas *legJ* is upregulated during the desiccation phase of cotyledon development, *legL* is regulated more like a "normal" storage protein gene with maximal levels of its mRNA present at 16−20 daf (i.e. the latter period of cotyledon expansion) and declining levels thereafter. The increase in the expression of *legJ* subfamily genes observed at 22−24 daf (THOMPSON et al. 1989) is due to *legJ* alone. The basis of this differential expression may become more clear if the sequence of *legL* is determined; it is, however, noteworthy that the *legJ* gene lacks the "core enhancer" sequence GCCACCTC in its 5′ flanking region; this sequence is present in all the pea *legA* family genes, and in the pea vicilin family genes sequenced, and homologous genes in other species as well as other highly expresed seed protein genes such as *Phaseolus vulgaris* lectin (GATEHOUSE et al. 1986), and may have an important influence on developmental control of expression.

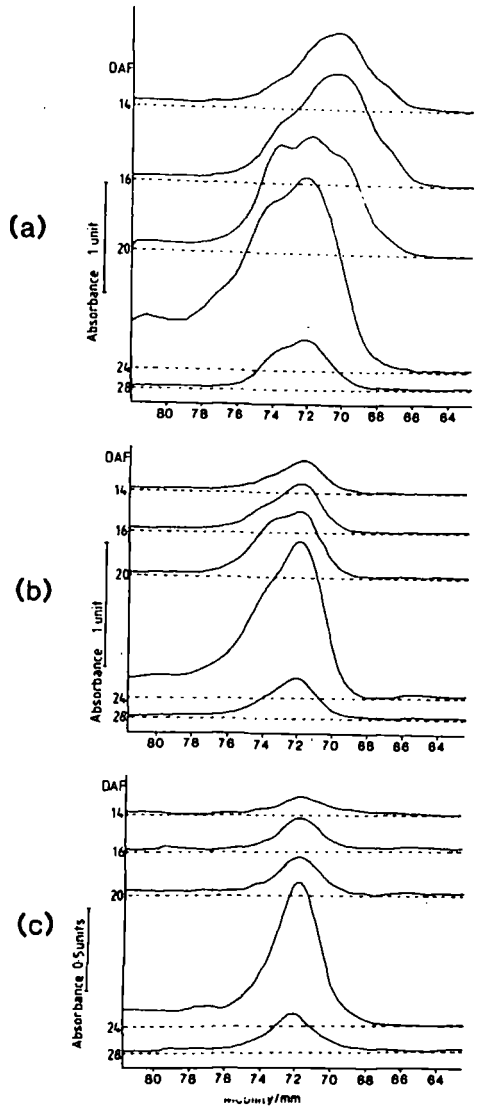


Fig. 3.B. *Densitometric analysis of autoradiographs shown in fig. 3A.* Individual tracks of autoradiographs shown in fig. 3A were scanned by laser densitometer LKB Ultroscan). Only tracks containing RNA treated to remove poly(A) tails ("−" tracks) were scanned. Numbers on vertical axes indicate developmental stage (DAF) of RNA samples. Dotted lines indicate the baseline for each scan. Scans are from autoradiographs probed as follows: (a), *legK* cDNA (detects expression of all genes in the *legJ* subfamily); (b), *legJ* oligo1; (c), *legJ* oligo2.

The relative strengths of the hybridisation signals of *legJ* oligo2 to the 72 mm band over RNA samples from different stages of seed development, and *legJ* oligo1 to the 72 mm and 74 mm bands were not significantly different to each other, nor to the whole coding sequence probe (*legk* cDNA) to the 72 and 74 mm bands. These data show that the 72 and 74 mm bands represent RNA species that are products of *legJ*, and clearly establish *legJ* as the

Biochem. J. (1988) 251, 717–726 (Printed in Great Britain)

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-^{35}S]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 NNCALN 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 μCi used/0.2–0.5 μg 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 × SSC (1 × SSC is 0.15 M-NaCl/0.015 M-sodium citrate buffer, pH 7.2)/2 × Denhardt's solution (1 × Denhardt's solution is 0.02% Ficoll/0.02% bovine serum albumin/0.02% polyvinylpyrrolidone)/denatured herring sperm DNA (100 μg/ml), at 65 °C as described in [20]; subsequent washes were to a hybridization stringency of 0.1 × 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 × SSC, 2 × Denhardt's solution/denatured herring sperm DNA (200 μg/ml)/50% (v/v) formamide, at 42 °C [22]; subsequent washes were to a hybridization stringency of 0.1 × 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) Ultroscan XL densitometer.

S1 mapping. S1 mapping was carried out as described by Favaloro *et al.* [23]. Each assay mixture contained 5 μg 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 μg (approx. 2×10^6 c.p.m.) of DNA probe, 5′ end-labelled [25] with [γ-^{32}P]-ATP (6000 Ci/mmol; 50 μCi used/0.2–0.5 μg 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 C_{18} 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 diaminobenzoyl 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

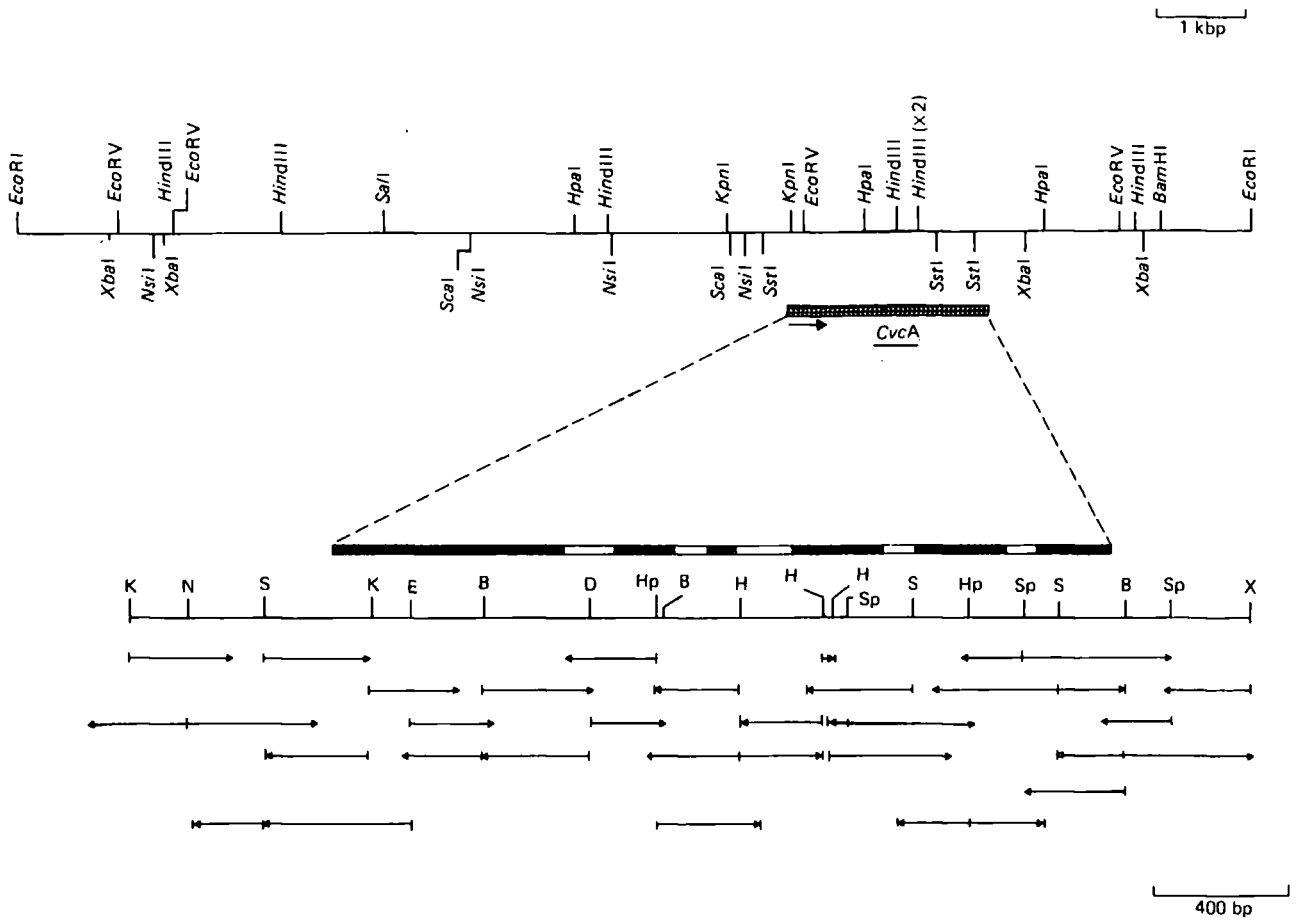


Fig. 1. Restriction map of the clone pJC4-100 containing *cvcA*, and sequencing map of *cvcA*

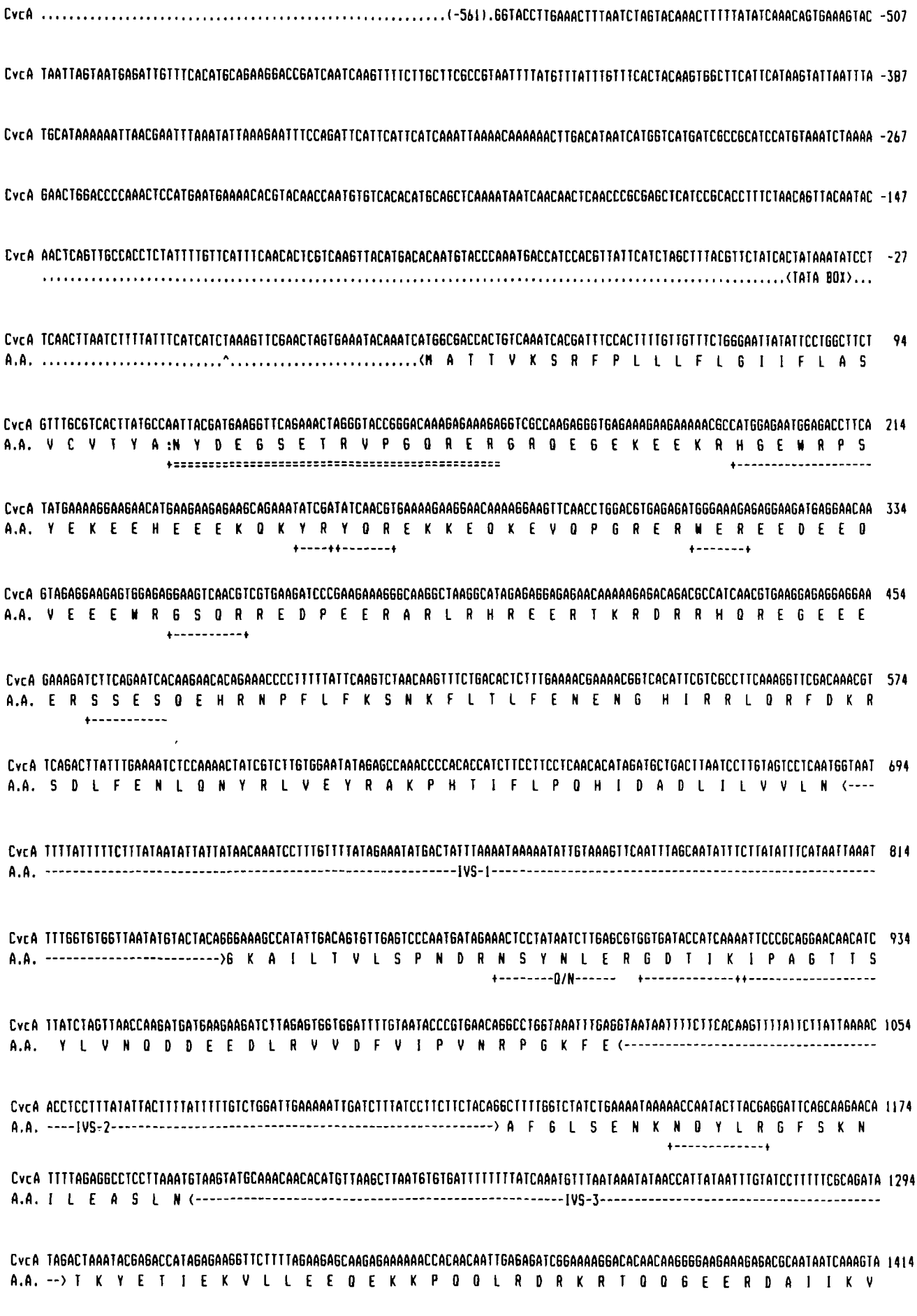
Key to restriction site symbols on sequencing map; B, *Bgl*II; D, *Dra*I; E, *Eco*RV; H, *Hin*dIII; Hp, *Hpa*I; K, *Kpn*I (= *Asp718*I; N, *Nsi*I; S, *Sst*I; Sp, *Ssp*I; X, *Xba*I.

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 *Asp718*I 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

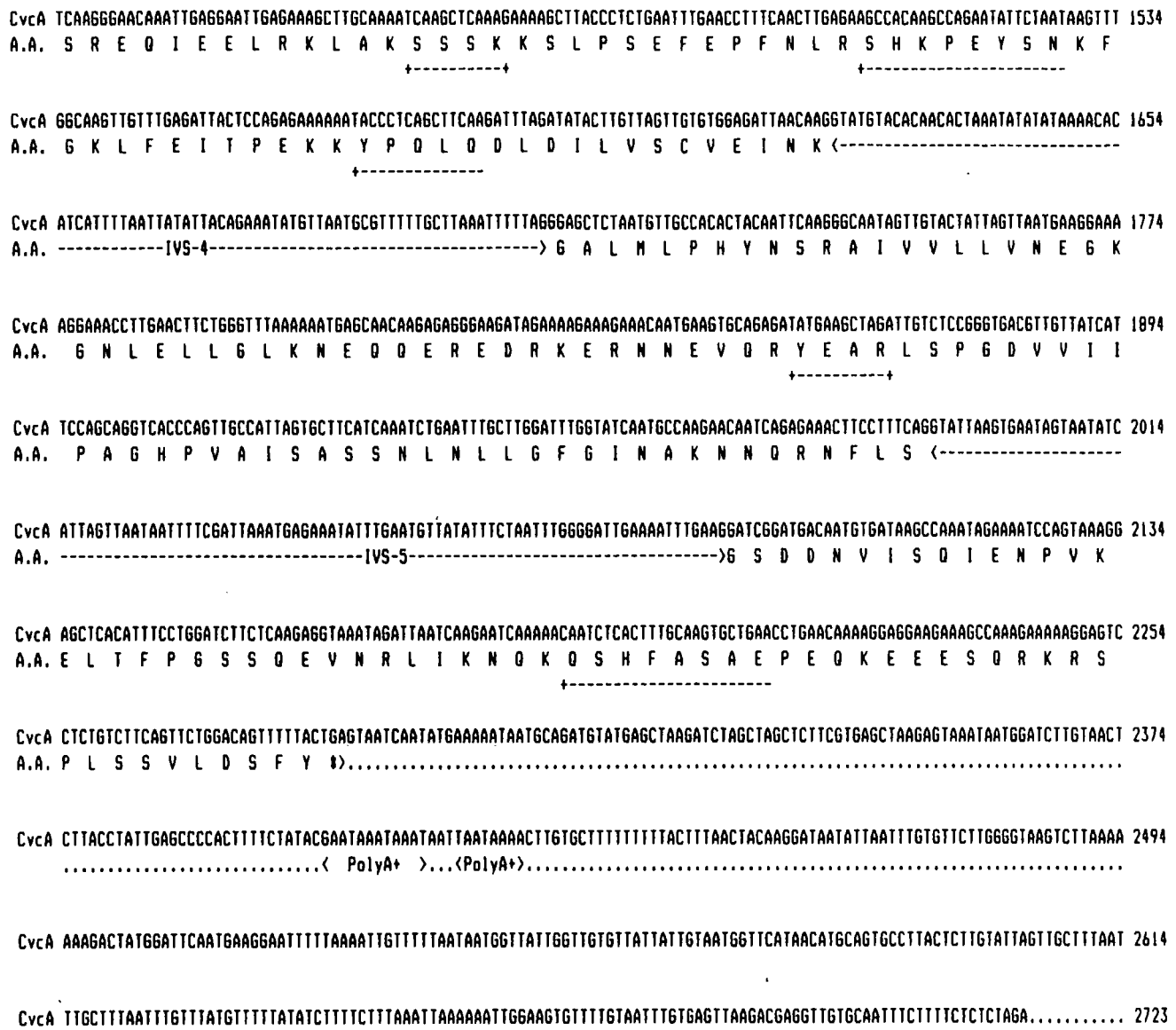


Fig. 2. Sequence of gene *cvcA* ('CvcA'), with the predicted sequence of the convicilin precursor polypeptide ('A.A.')

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

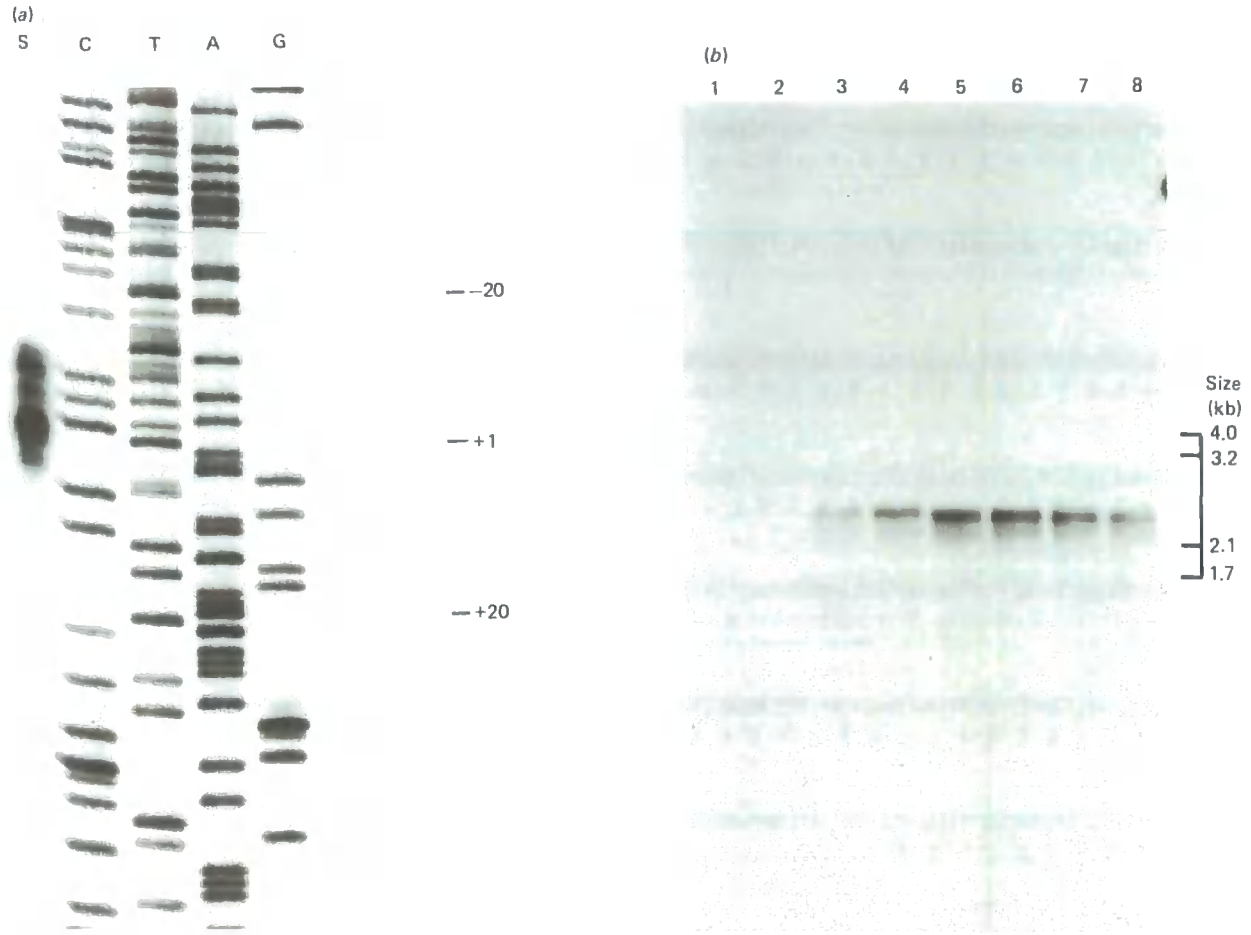


Fig. 3. Expression of convicilin gene *cvcA*

(*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 μg 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 α:β 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

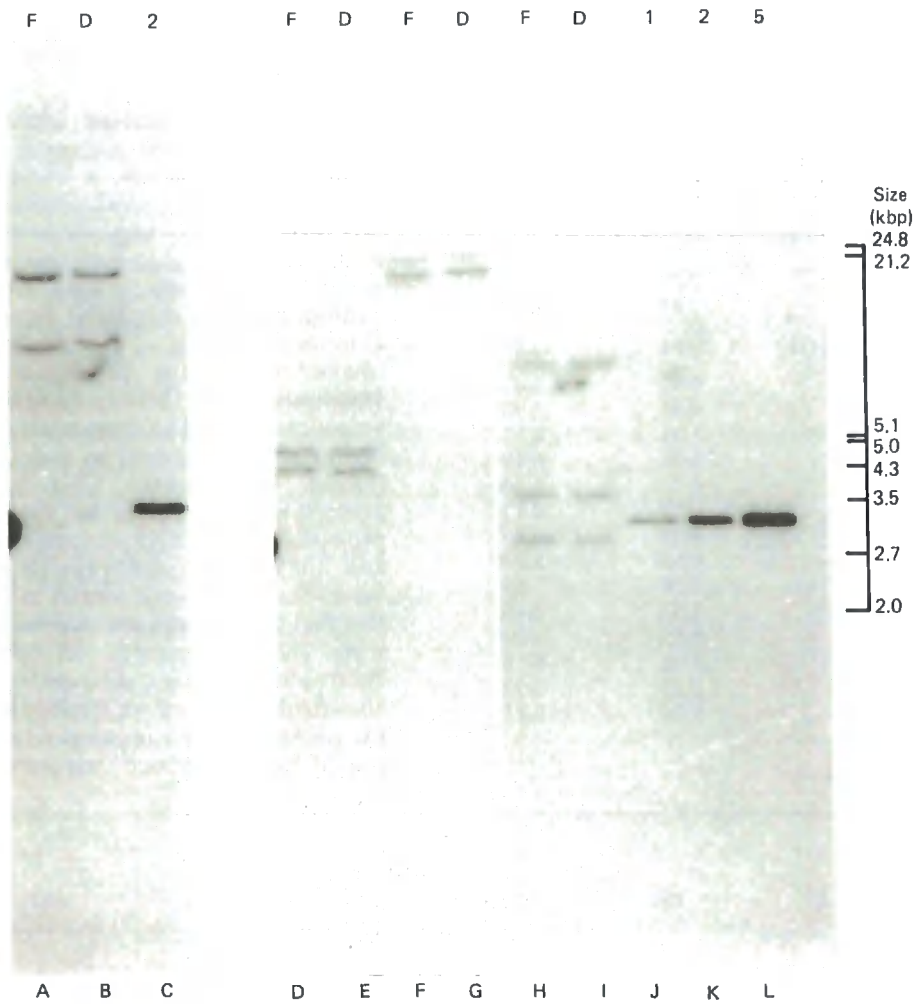


Fig. 4. Southern blot showing hybridization of *Sst*I–*Bgl*II probe (bases −176 to 462) from *cvcA* to restriction digests of genomic DNA from lines Feltham First (F) and Dark Skinned Perfection (D)

A 10 μg portion of DNA was loaded per track on the original gel electrophoresis. Restriction enzymes used were as follows: A and B, *Eco*RI; D and E, *Bgl*II; F and G, *Bam*HI; H and I *Eco*RV. 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 α:β and β:γ 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 that 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		2.39	2.55
S	40		7.37	6.39
E	80	113	20.81	22.08
Q	33			
P	25		4.60	5.47
G	27		4.97	5.90
A	18		3.31	4.23
C	1		0.17	0.13
V	27		4.97	4.46
M	1		0.17	0.13
I	24		4.42	3.85
L	49		9.02	8.71
Y	15		2.76	2.59
F	20		3.68	3.30
W	3		0.55	ND†
K	43		7.92	8.18
H	12		2.21	2.22
R	53		9.76	8.15

* 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.

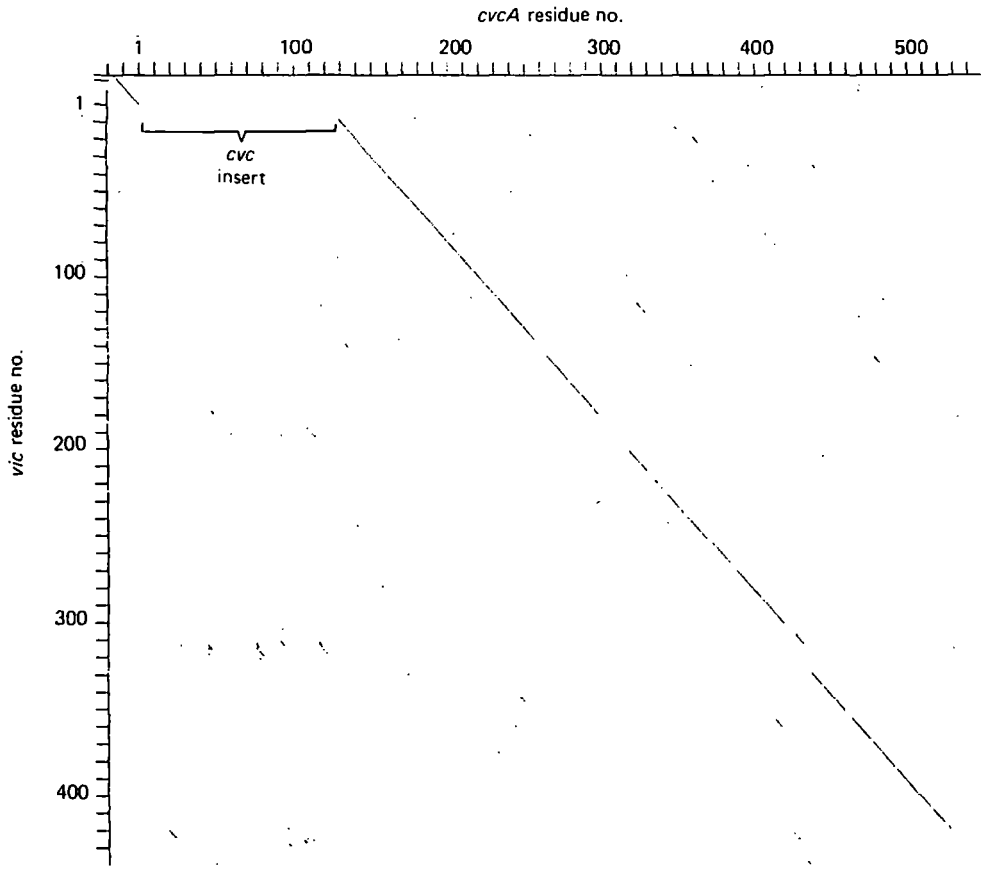


Fig. 5. Dot-matrix comparison of the amino acid sequences of vicilin (from pAD 2.1 plus *vicB*) and convicilin

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-

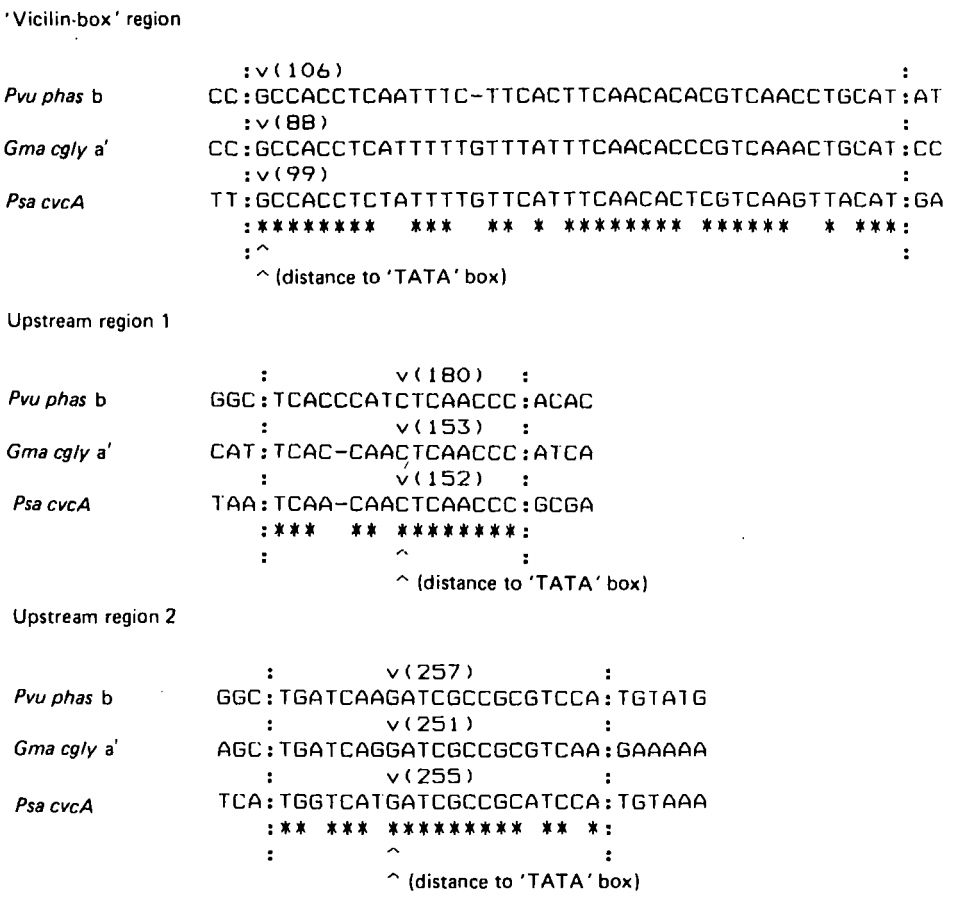


Fig. 6. Putative enhancer sequences in the 5′ flanking regions of *cvcA*

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 Agrobiological 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

Received 7 May 1987/14 October 1987; accepted 21 December 1987