Comparative studies of the nucleotide sequences of pea seed storage protein genes

M. D. Levasseur

TO MY MOTHER

Comparative studies of the nucleotide sequences of pea seed storage protein genes

by

M. D. Levasseur BSc. Dunelm

A thesis submitted in accordance with the requirements
for the Degree of Doctor of Philosophy
in the University of Durham

Department of Botany April 1988

Abstract

Nucleotide sequence data from several pea (*Pisum sativum* L.) seed storage protein genes was obtained. Of two legumin genes sequenced, one was shown to be a pseudogene, apparently once coding for a polypeptide belonging to the 'major' legumin class, whilst the other was shown to be a functional gene coding for a polypeptide of the 'minor' legumin class. Sequence data was also obtained for two vicilin genes. Complete sequencing of one revealed it to be truncated by sequence of unknown origin at its 3' end, whilst partial sequence for the other suggested the presence of a stop codon in the coding region. These findings implied that both vicilin genes are no longer functional.

Additionally, various comparisons of nucleic acid and amino acid sequence data were made between these genes and also with other legume seed storage protein genes. Results showed these genes conform with the major structural features of eukaryotic genes, and also revealed the presence of potential tissue-specific regulatory elements in the 5' flanking regions of the genes. Dendrograms for legume 11S and 7S classes of globulin seed storage protein genes clearly supported the model theory of each class having arisen by successive duplications from a common ancestral gene.

List of Contents

LIST OF FIGURES

Fig. Page

LIST OF TABLES

Table

Acknowledgements

I would like to thank Professor D. Boulter and Dr J.A. Gatehouse for all the valuable assistance given during their supervision of this project.

My thanks also go to Mr D. Bown for excellent technical assistance, to Dr M. Evans, Dr R.N. Waterhouse, Mr D. Bown and Mr A. Sherman for many helpful discussions, to Mr P. Sidney for photographic work, and to Mrs K.E. Leck and Mrs K.L.M. Nattrass for typing of the manuscript.

This project was funded by an S.E.R.C. Research Studentship grant.

Memorandum

Parts of the work described in this thesis have previously been presented in the following publications

Bown D., Levasseur M., Croy R.R.D., Boulter D. and Gatehouse J.A. 1985; Sequence of a pseudogene in the legumin gene family of pea (*Pisum sativum* L.) N.A.R. 13 4527-4538.

Gatehouse J.A., Bown D., Gilroy J., Levasseur M.D., Castleton J. and Ellis T.H.N. 1988; Two genes encoding 'minor' legumin polypeptides in pea (*Pisum sativum* L.) Biochem. J. 250 15-24.

Abbreviations

Abbreviations were those used according to the Biochemical Society Instructions to Authors, Biochem. J. 1982 209 1-27, with any additions given below.

The one letter notation for amino acids is given in Biochem. J. 1969 113 1-4.

Kb	=	Kilobase (pairs)
bp	=	base pairs
Et Br	=	Ethidium Bromide
mRNA	=	messenger RNA
tRNA	=	transfer RNA
cDNA	=	complementary DNA
dCTP	=	deoxycytidine 5'-triphosphate
d.a.f.	=	days after flowing
LMP Agarose	=	Low Melting Point Agarose
SDS	=	Sodium Dodecyl Sulphate
SSC	=	Saline Sodium Citrate
Y	=	Unspecified pyrimidine nucleotide
X	=	Unspecified purine nucleotide
N	=	Unspecified nucleotide
IEF	=	Isoelectric focussing
SDS - PAGE	=	Sodium Dodecyl Sulphate - Polyacrylamide gel electrophoresis
X - gal	=	5-Bromo-4-chloro-3-indolyl- β -D-galactopyranoside
IPTG	=	Isopropyl- β -D-thiogalactopyranoside
ug	=	micrograms
ul	=	microlitres
Ci	=	Curies
uCi	=	micro Curies

CHAPTER 1 INTRODUCTION

1.1 Eukaryotic Genes

1.1.1. The Eukaryotic Genome

Eukaryotic and prokaryotic genomes show major differences in the organisation of their DNA, the main difference being that eukaryotic genomes in general contain much more repeated sequence DNA and non-coding DNA than do prokaryotic genomes.

The DNA content of the eukaryotic haploid genome is known as the C value, and ranges from a mere 10^4 bp for a mycoplasma, right up to as much as approx. 10^{11} bp from some plants and amphibians (Lewin, 1983). However, from the fact that DNA content within genera can differ by up to 10 fold (Rees and Hazarika, 1976, Jones and Brown, 1976), it seems evident that genome size is not directly linked to the complexity of the organism. Indeed, studies using reassociation kinetics reveal 3 components within the DNA, which can be separated according to speed of hydridisation. These correspond to non repetitive, middle repetitive and highly repetitive sequences (Britten and Kohn, 1968). Of these components, the non repetitive element is the smallest e.g. in pea it has been shown to represent only 19% of the genome, with only a small portion of this actually thought to represent protein coding sequence (Murray et al., 1981). In mammalian globin genes, only 8% of the DNA within the gene clusters is thought to represent coding sequence, 8% representing introns and the remaining 84% comprising a mix of single copy and repeat elements in intergenic arrangement (Jeffreys, 1981).

The concept of such excessive DNA being termed 'junk' DNA was suggested by Orgel and Crick, (1980) implying it had no function. Several pieces of evidence and lines of thought since then however seem to refute this suggestion.

Firstly, comparison of the entire β -globin gene cluster between man, gorilla and baboon (Barrie, 1981) shows it to be highly conserved in size and composition between the three species, implying a large degree of functional constraint, and prompting Bennett, (1982) to accord the status of the entire cluster as a sort of 'supergene' of single co-adapted function, his reasoning being that conservation of lengths of non coding DNA was acting to maintain spatial arrangements of the genes comprising the cluster on the chromosome.

An additional consideration, related to the packaging of DNA into nucleosomes (Kornberg, 1974, 1977,) within chromatin filaments (reviewed by Igo-Kemenes, et al., 1982, and McGhee and Felsenfield, 1980), is the notion of specific nucleosome phasing (ie specific placement of histone octamers with respect to DNA sequences,(Kornberg, 1981) to potentially allow exposure of more desirable DNA sequences by location in the more accessible linker regions between the cores of adjacent nucleosomes. Noncoding and repeat sequences could well be considered to play a very important role in maintaining this phasing.

Functional or not, it seems safe to say that the presence of such large amounts of excess DNA has played, and will continue to play an important role in genome evolution. In the Graminae, 75% of total DNA consists of repeated sequences (Rimpau, et al., 1978, 1980) and these can be classified into hybridisation families (where sequences belonging to the same family are sufficiently homologous to form stable duplexes under certain conditions - Flavell, 1982). Repeats of some families are clustered in long tandem arrays occurring on several or all chromosomes (strong evidence that a translocation mechanism acts - Bedbrook, 1980a, 1980b, Gerlack and Peacock, 1980, Dennis, et al., 1980a, 1980b, John and Miklos, 1979)

However most families of repeats are organised in a much more complex way (Flavell, 1980, Flavell et al., 1981, Wensking, 1979). Some of their members are interspersed with short non-repeat DNA segments and/or unrelated repeated sequences. Studies of this kind have led to the conclusion that the mechanisms acting to give rise to such a genome composition are basically sequence amplification (by duplication), deletion, rearrangement and translocation, usually acting on small segments of DNA, acting in a stochastic fashion, with some members surviving under selection, and others being deleted for the same reason (Flavell, 1982). A computer simulation of genome evolution (Loomis and Gilpin, 1986) led to the conclusion that genome size only stablised when the amount of dispensable sequences had increased to the point that most deletions did not affect vital genes, and in such genomes the number of copies of specific genes fluctuated, generating small multigene families. These results reflect the situation in the real eukaryotic genome, and imply that when the genome size is not critical to survival - as is the case in most eukaryotes the genome carries vestigial sequences which may or may not be functional, and that the occurrence of multigene families may often happen by chance.

The most fitting concept to apply to the eukaryotic genome is that it is in an ever continuing state of flux and that the constituent DNA is 'turning over' by continuous cycles of amplification and deletion (Flavell, 1980, 1981, Thompson and Murray, 1980, Dover, 1981, Dover et al., 1982).

1.1.2. The Structure of Eukaryotic Genes

Eukaryotic genes have been shown to contain several consistent features. Current understanding of this subject has largely been based on comparisons between animal gene sequences. However, as the volume of data on plant genes has increased, it has become apparent that plant and animal genes share many of these common features.

1.1.2.1 Introns

Introns are sequences of DNA present in the gene which interrupt the transcribed, non-translated sequence. They are processed out of the initial transcript from the gene, and as a result do not appear in mature messenger RNA. They occur in between the coding regions, or exons, of the gene to form a complex mosaic and can vary in length from 50 to 20,000 bp. (Gilbert, 1985). Since introns are often much larger than exons (an average coding length of 40 - 50 amino acids is most common), genes can consequently be as large as 200 kilobases (Gilbert, 1985).

Introns are common in higher eukaryotic genes, but some plant genes are notably 'intronless', including all zein genes so far characterized (Hu et al., 1982, Wienand et al., 1981), 2 soybean 15 kd storage protein genes of unknown function (Fischer and Goldberg, 1982), and all lectin genes so far examined (Gatehouse et al., 1987).

With the advent of nucleic acid sequencing the exact position, length and sequence of introns can now be determined. Most notable is the conservation of sequence at 5' and 3' splice sites (Breathnach et al., 1978) common to both animal and plant genes. A more recently observed consensus sequence also thought to play a role in splicing out of the introns from nascent mRNA is the branch point (Wallace and Edmonds, 1983, Keller and Noon, 1984, Brown, 1986). The consensus sequence differs slightly between animal and plant introns, and in plants, is of the form T(Pu) T(Pu) C/T T Pu A Py (Brown, 1986). It is thought to act as the recognition site for formation of a lariat RNA - an intermediate in the splicing event - and operates in conjunction with small nuclear ribonucleoprotein particles (sn RNPs) in achieving this (Brown et al., 1986, Gerke and Steitz, 1986, Sharp, 1987). Conservation at the intron/exon boundaries also plays a contributory role in splicing, possibly involving the action of a single, common enzyme (Breathnach,et al., 1978). Any role for potential secondary structure formation within introns has yet to be ascertained, but studies by Thimmappaya and Schenk, 1979, using viable mutants of SV40 lacking large regions of the large T intron seem to imply an absence of function for any potential secondary structure.

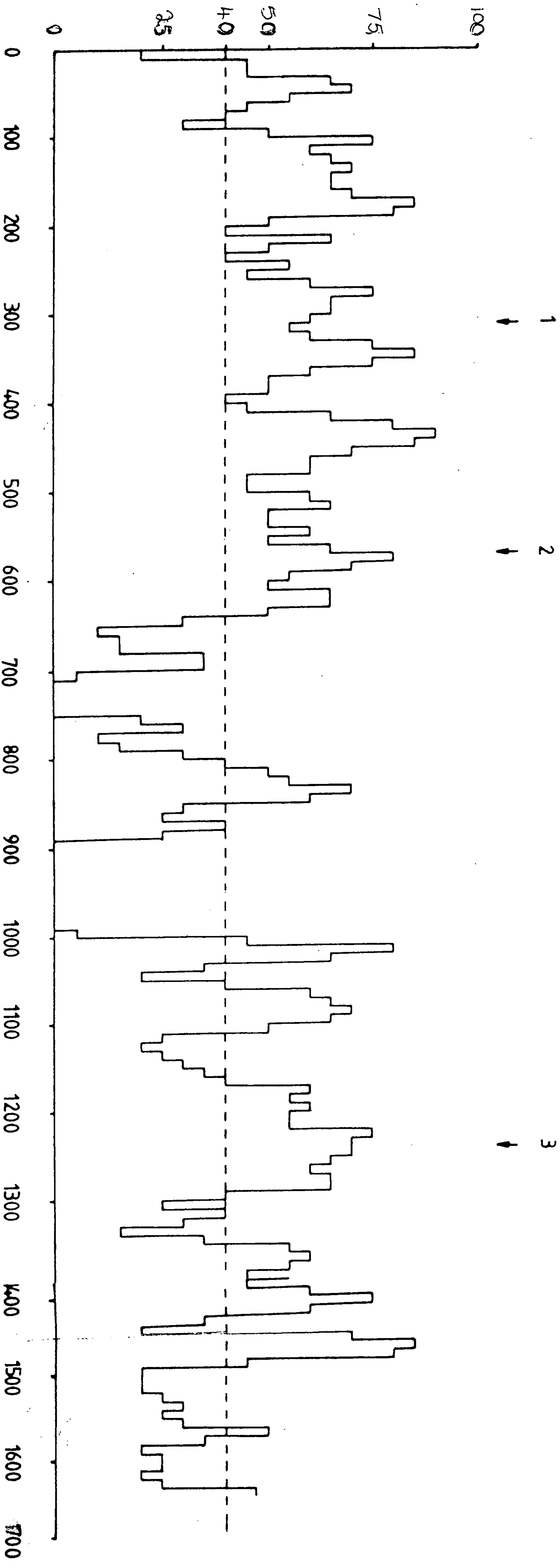
Intron function remains unclear. As stated, they are usually situated between exons, however they have been found to occur in 5' flanking sequence (e.g. Ovalbumin (Breathnach et al., 1978)), but not, as yet, in 3' flanking sequence, implying there is no selective advantage in splicing the 3' end of a messenger RNA to bring it closer to the termination codon (Breathnach et al., 1978). A further factor to consider is the differences observed in intron number between groups of genes e.g. a maximum if 5 in pea seed storage proteins (see section 4) compared to 33 in vitellogenin genes of amphibians (Wahli et al., 1980). It is not known why this should be so, although it may tie in with the concept of introns acting as delineators between functional domains coded by exons, and hence facilitating formation of new proteins or improvement of existing ones by exon shuffling (see Section 4.1.2.1)

1.1.2.2. 5' Flanking Sequence

Gene expression is controlled at least partly at the level of transcription. Regulation mechanisms involve a variety of sequences in the DNA specifying levels of transcription and startsite of the message. Comparison of 5' flanking sequence between genes reveals consistent presence and position of such signals, termed promotors.

i) The TATA Box

The TATA or Goldberg - Hogness box is the best characterised of eukaryotic promotor elements (Proudfoot, 1979) and has been shown to play a role in transcription initiation (see below). It is the eukaryotic equivalent of the Pribnow box found in prokaryotes (Pribnow, 1975a, 1975 b, Schaller et al., 1975), and is consistently located 26-34 nucleotides (measured from the second T) upstream from the cap site (see below) (Messing, et al., 1983).


0
25
40
50
75
100
0
100
200
300
400
500
600
700
800
900
1000
1100
1200
1300
1400
1500
1600
1700
1
2
3

It functions by interacting with RNA Polymerase II (which transcribes the message) and ensures that transcription is initiated in the correct place (Breathnach et al., 1978). Evidence for this comes from deletions in the TATA region at E1A in Adenovirus (Osborne et al., 1982), showing that absence of the TATA element results in a 5 to 10 fold decrease of E1A mRNA, and also the appearance of a set of E1A messages of slightly different lengths due to relaxation of stringency in transcription initiation. A similar result was observed by Mathis and Chambon, 1981, with deletions in the TATA box of the SV40 Early region. Creation of a point mutation in the TATA box from conalbumin, fused to the sea urchin H2A histone gene (known to be efficiently transcribed in Xenopus oocytes) showed consistency of start site in transcription initiation between the normal TATA sequence and its mutation, but the level of transcription in the latter was reduced by 5 times (Grosschedl et al., 1981).

ii) The CAAT Box

Another sequence that may be involved in the regulation of transcription of some eukaryotic genes is the consensus sequence GGC/TCAATCT or 'CAAT-box' (Benoist et al., 1980) and is found 80-100 nucleotides upstream of the cap site (see below).

The role of the CAAT box in transcription regulation was suggested by deleting the CAAT box in globin genes in Mouse L cells (Dierks, et al., 1981) where transcription was consequently markedly reduced, and by similar work on the rabbit β -globin gene *in vivo* (Grosveld et al., 1981).

Occurrence of this sequence in the equivalent position in plant genes appears to be limited. Two zein genomic clones Z4 and Z7 (Hu et al., 1982, Kridl et al., 1982) have sequences with limited homology to the CAAT box, and 3 sequences upstream from the coding region in soybean leghaemoglobin genes have homology to the animal sequence (Brisson and Verma, 1982). Indeed, it seems that cereal genes (with the exception of the maize ADH-2 gene, Dennis et al., 1984) commonly display this feature, but it was not found in a manual survey of 15 plant genes scanned (Kreis M, 1986).

A possible counterpart of the CAAT box in plants was identified by Messing et al., 1983. They termed the possible regulatory sequence element the 'AGGA box', and it refers to the symmetry of adenines surrounding the trinucleotide G/T N G. Such a sequence was identified in the 5' flanking sequence of the Legumin A gene of *Pisum Sativium* (Lycett et al., 1984) but any functional significance of the element is yet to be elucidated.

iii) <u>Enhancers</u>

Enhancers were discovered in 1981 (Gruss et al., 1981) and defined as potent gene activators of transcription in both viral and cellular genes. They have been identified as sequences of DNA of varying length, and their effectiveness appears largely unaffected by orientation and positioning with respect to the gene and can act up to 10kb away.

The first discovered enhancer was in SV40, and took the form of a 72 bp tandem repeat, 100 bp upstream from the Cap site of early viral genes and was found to act as a cis regulatory element (Benoist and Chambon, 1980). Deletion of this region was shown to decrease early viral gene expression by up to 100 times (Gruss et al., 1981). Similar decrease in expression levels was obtained with deletions 138 to 194 bp upstream of the genes involved in amino acid biosynthesis in yeast (Donahue et al., 1983). Further evidence for the potent action of enhancers came from the findings that the SV40 enhancer plus analogues isolated from other animal viruses can function not only when linked to their natural genes but also in association with heterologous genes (Banerji et al., 1981).

Several incidences of enhancers being found in plant genes have now been documented. The first discovered was an upstream element of a Rubisco small subunit (rbcs) gene in pea which could, in both orientations, confer light-inducible expression on a heterologous promoter/gene (Timko et al., 1985). Additionally, the existence of a 247 bp element from the light harvesting chlorophyll protein (Lhcp) gene in pea was reported which appears to act not only as a light - inducible enhancer, but also as a tissue-specific 'silencer' (Simpson et al., 1986). Additionally, the presence of a short 49 bp sequence element with enhancer like properties was discovered upstream of the α submit gene of β -conglycinin in soybean (Chen et al., 1986), and appears to play an important role in controlling the expression of this embryo tissue-specific gene.

The way in which enhancers work is still not entirely clear. One theory is that, since viral enhancers tend to be in regions of increased DNase sensitivity, this may cause an alteration in the chromatin structure, in turn possibly causing increased transcriptional activity (Banerji et al., 1981). More evidence exists now though to suggest that the specific sequences present in the cis-acting enhancer element provide a binding site for some trans acting factor, consequently enhancing transcription possibly by an interaction with RNA Polymerase II (Khoury and Gruss, 1983, Dynan and Tjian, 1985, Voss et al., 1986). Such a nuclear binding factor has been observed in zein genes from maize (Maier et al., 1987). Specific binding was involved at a 22 nucleotide stretch, which includes 14 bp out of a 15 bp sequence conserved in all zein genes (Brown et al., 1986). Similar results were obtained with the B1 hordein gene in oats (Kreis et al., 1986). 6 DNA binding proteins were identified that were thought to recognise and interact with putatitive regulatory sequences identified not only in the B1 hordein gene, but also in other prolamin genes (zeins in maize, α gliadins in wheat). Whatever their mode of action, cis-acting enhancers are thought to play a major role in the specific expression of vicilin and legumin genes in pea, giving characteristic differences in mRNA timing and patterns of accumulation within each (Higgins, 1984).

Some light is shed on possible origins of enhancers in cellular genes when one considers the observation that the organisation and structure of the enhancer element present in the human metallothionein IIA gene is very similar to that of the SV40 enhancer, suggesting a possible evolutionary link between viral enhancers and cellular upstream regulatory elements. Indeed, in support of such a theory, a region homology to the SV40 core enhancer sequence has been observed in the 5' flanking sequence of the legumin A gene (*Leg* A) in pea (Lycett et al., 1985), along with the previous observation of another short sequence also in LegA showing 80% homology to the adenovirus enhancer core element (Hearing and Schenk 1983, Lycett at al., 1984).

iv) The Cap Site

5' terminal capping of messenger RNA with 7 methyl guanosine to form a cap structure of m^7G(5')ppp(5')N was originally discovered in 1974 (Reddy et al., 1974). It seems the 5' terminal cap structure in eukaryotic mRNAs is different to that in prokaryotes, in that it shows resistance to phosphorylation following treatment with alkaline phosphatase (Rottman et al., 1974). Capping is present almost universally in eukaryotic mRNAs, the only reported exceptions being the poly(A)-containing mRNAs isolated from HeLa cell mitochondria (Grohmann et al., 1978, Taylor and Dubin, 1975).

The capping reaction occurs at the 5' triphosphate ends of nascent pre-mRNAs shortly after initiation of transcription and is catalyzed by guanylyl and methyltransferases (Shatkin, 1985). During processing of nuclear transcripts, the cap is retained, serving as a stabilizing element both on pre-mRNAs in the nucleus. (Green et al., 1983) and mRNA in the cytoplasm (Furuichi et al., 1977). The cap's importance in ensuring normal messenger RNA splicing is demonstrated by the use of cap molecule analogues, causing a marked inhibition of splicing (Konarska et al., 1984). The presence of the cap also appears to markedly enhance translation by promoting initiation complex formation. Use of analogues has shown inhibition of attachment of the

40S ribosomal subunits to capped mRNAs, implying that recognition of the capped end of mRNA by specific protein(s) is important for gene expression (Shatkin, 1985).

1.1.2.3 3' Flanking Sequence

The major feature of importance in the 3' flanking sequence is the consensus sequence AATAAA. Commonly known as the Polyadenylation signal, this sequence was discovered in the mRNA's of several unrelated mammalian genes (Proudfoot and Brownlee, 1976). Since then, it has become apparent that such a sequence is present in all eukaryotic genes apart from histones and yeast RNA Polymerase II type genes (Proudfoot, 1982).

Evidence for the sequence's roles as a polyadenylation signal came from work on SV40. Mutants lacking the late gene AATAAA, but with it present further downstream showed no wild type 3' terminus, but one with a poly(A) tail 15 bases away from the AATAAA signal (Fitzgerald and Schenk, 1981). Generally the sequence is found 11 to 30 bases away from the poly(A) tail (Gil and Proudfoot, 1982), however, it seems in plant genes, it tends to differ slightly in its position in being further away from the poly(A) tail, and nearer the stop codon.

It is now thought that the sequence does not alone comprise the entire polyadenylation signal, but acts along with another factor(s) to cause the poly (A) tailing (Gil and Proudfoot, 1982). Such additional elements are thought to be possibly either specific secondary structures in the message, or other significant linear sequences (Manley, 1983). What is clear is that polyadenylation occurs at a single downstream site from AATAAA, except in the case of bovine prolactin (Sasavage et al., 1982) mouse ribosomal protein L30 (Weidemann and Perry, 1984), and pea seed lectin (Gatehouse, 1986) mRNAs, where in each case there is a single signal site, but poly(A) is at several sites within a 12-14 nucleotide region.

The following hexameric sequences have also been found from a search of the literature and data banks (Birnsteil et al., 1985) to occur naturally at the appropriate distance from poly (A) addition sites; AATTAA, AATACA, AATAAC, CATAAA (Wickens and Stephenson, 1984) ATTAAA, AGTAAA, TATAAG and AATATA, and it is plausible therefore, that they also may function as polyadenylation signals.

It has also been observed that many plant genes have more than one potential polyadenylation signal eg. soybean leghaemoglobin has two (Messing, 1983) or pea legumin gene A, which appears to have 3 double overlapping signals, with cDNA evidence strongly suggesting exclusive use only of the second of these signals (Lycett et al., 1984). This legumin gene also offers a site of potential secondary structure downstream of the signals which could play a role in determining the choice of polyadenylation site (Lycett et al., 1984). Data so far seems to suggest that multiple polyadenylation sites occur more in plant genes than animal, where such a situation has rarely been observed (Early et al., 1980, Setzer et al., 1980, Tosi et al., 1981).

1.1.3 Organisation and Evolution of Eukaryotic Genes

It has become increasingly clear over the past few years that a large range of transciptional products are encoded by more than just one gene, and indeed, an extreme example is the case of 5S ribosomal RNA in *Xenopus laevis* where a gene copy number of 24,000 has been reported (Brown et al., 1971). Evidence gained so far points to a correlation between proteins present at high levels within a cell, and its corresponding gene(s) being present in multiple copy number, thus allowing high levels of mRNA to be transcribed and hence high quantities of the protein to be produced. Examples of such functional families are the genes coding for chicken oviduct proteins (Axel et al., 1976), or contractile protein genes (rev. by Buckingham and Minty, 1983).

Organisation of the genes in these multigene families tends to be in 2 main patterns; arrays, of varying lengths, of tandem repeats interspersed with repetitive DNA, to form a cluster, with all the genes linked closely on one chromosome, or representation within clusters either widely spaced on the chromosome or even on entirely different chromosomes. (Jeffreys and Harris, 1982). An example of both these situations occurs in the globin gene family in mammals, with two unlinked gene clusters coding for the α and β globins. In the human β -globin gene cluster, there are 5 active genes $-\varepsilon - {}^{G}\gamma - {}^{A}\gamma - \partial - \beta$ plus two pseudogenes (see sections 1.1.4) and extensive tracts of non coding DNA between these genes (Efstratiadis et al., 1980). Nucleotide sequencing of an 11 kb fragment encompassing the two foetal genes, ${}^{G}\gamma$ and ${}^{A}\gamma$, gives clear evidence that duplication has been the major active mechanism in creating the cluster (Smithies et al., 1981). Indeed, further analysis of this segment showed the presence of short direct repeats at each end of the genes ie.r-$^{G}\gamma$ -r-$^{A}\gamma$ -r (Shen et al., 1981), suggesting a simple model of mispairing and unequal crossing over at meiosis between these repeat elements in an ancestral r- γ -r DNA segment. Linking this to the fact that short repetitive elements are extremely prevalent in the mammalian genome suggests that large block duplications could occur quite commonly (Jeffreys and Harris, 1982). The presence of dispersed repeat elements could also lead to the excision of genes via intrachromosomal recombination, and resulting excised circles could then integrate elsewhere in the genome by homologous recombination with another member of the repeat family (Jeffreys and Harris, 1982).

The fact that members of a multigene family often maintain remarkable levels of homology, even when members are unlinked (e.g. primate β - and α -globin clusters), led to the suggestion that such 'homogenisation' may have occurred via a recent gene conversion event (Slightom et al., 1980, Shen et al., 1981). A similar situation appears to exist in the ribosomal RNA gene family of Xenopus (Dover and Coen, 1981).

To summarise, it seems the single most important mechanism occurring to produce multigene families is that of duplication - either by tandem gene duplication, chromosome duplication, or genome duplication (polyploidisation) (Li, 1983). Subsequent to this, members of the family could then move to other locations in the genome (possibly by the above mentioned process of repeat sequence recombination), where they then may proceed to evolve independently, or be subjected to functional constraint or interchromosomal gene conversion.

1.1.4 Evolution of Pseudogenes

The term 'pseudogene' was first proposed to described genes within the 5S RNA cluster of *Xenopus laevis* clearly related to the normal 5S genes, but incapable of producing functional 5S rNA copies (Jacq et al., 1977). Basically, pseudogenes are non-functional counterparts of once active genes, which, due to the gradual accumulation of mutations (such as point mutations, deletions and insertions) have lost the potential to be transcribed, or for their mRNA to be translated (Lewin, 1983).

There appears to be two general categories to classify pseudogenes (Vanin, 1984). Firstly there are those which retain the intron/exon arrangement of their productive counterparts e.g. cases within the globin gene family (Lacy and Maniatis, 1980, Proudfoot and Maniatis, 1980). The second category includes those which lack the intervening sequences found in their active counterparts, this type having been labelled 'processed pseudogenes' (Vanin, 1984). Examples of this type are becoming more numerous, with the first found being in mouse α -globin genes (Nishioka et al., 1980 Vanin et al., 1980). Aside from various genetic lesions present in this type of pseudogene, other interesting factors have emerged e.g. homology between them and their productive counterparts ceases at the points corresponding to initiation and termination of transcription. Also, immediately 3' to the point at which this homology breaks down, there is often found an oligo (A) tract anything from 11 to 38 nucleotides in length. All this strongly implies that the origin of processed pseudogenes must have been by a reverse transcription event from the original messenger RNA from the active gene (Vanin, 1984). Further support for this theory comes from two sources. Firstly, the observation in the mouse Ψα 3 processed pseudogene that it is flanked by sequences homologous to retroviral-like mouse intracisternal A particle RNA, suggesting a transposition (Lueders et al., 1982). Indeed the presence of direct repeats on the flanks of these genes has been found in many other instances (rev. by Vanin, 1984). Linked to this evidence is the fact that processed pseudogenes and their active counterparts are not found on the same chromosome, whereas pseudogenes which retain the intron/exon structure of the active counterparts, appear to occur in tandem arrays (Vanin, 1984)

In mammalian genomes, cases of processed pseudogenes are now almost as numerous as those for non-processed, however no cases of the processed type have been found outside mammalian systems. (Vanin, 1984)

Plant examples of pseudogenes have also been found e.g. the soybean leghaemoglobin pseudogenes (Brisson and Verma, 1982) and in maize zeins (Spena et al., 1983).

It would seem justifiable to assume a possible dual stage in non-processed gene evolution; Firstly, after a gene duplication event, a slow rate of mutation would ensue, according to the degree of selection pressure, until inactivation of the gene occurred. At this point, any functional constraint would cease to operate, and the mutation rate would accelerate (Little, 1982). In the case of processed pseudogenes, if the reverse transcription theory is to be believed, there would be no inactivation as such, since the reverse transcript lacking introns would never be active in the first place and consequently mutation would occur at a high rate from this point.

Postulating a functional role for pseudogenes in the genome might seem inappropriate but one suggestion has been that their presence and position is maintained due to the importance of not disrupting chromatin structure (Jeffreys, 1981). However, one might prefer to regard them as merely a component of an everchanging genome, only disappearing at all when their presence becomes deleterious to the cell phenotype.

1.1.5. Molecular Evolution

The concept of molecular clocks to describe the rate of evolutionary change in related molecules is now well established, and was first proposed after evolutionary studies on globin genes (Zuckerlandl and Pauling, 1965, Jukes and Cantor, 1969). In order to construct a molecular clock, 3 basic requirements must be met; Firstly, a measurable event must be involved. Secondly, such an event must occur regularly (either metronomic, or stochastic, the latter being more likely) (Fitch, 1976). Thirdly, the clock must be calibrated by some external event of known date e.g. the divergence of two species at a time known from the fossil record (Ferguson 1980). Consequently, in recent years, with the increasing availability of nucleic acid sequence data, molecular clocks have been constructed using several classes of genes e.g. chicken insulin genes (Perler et al., 1980), globin genes (Lacy and Maniatis, 1980) and actin genes (Shah et al., 1983). The use of nucleic acid sequence data instead of amino acid sequence in comparisons is obviously superior, in that the former reflects silent mutations (those not causing a change in an amino acid) as well as replacement mutations (those which do cause a change), whereas, due to the redundant nature of the amino acid code, comparisons between amino acid sequences cannot do this.

However, amino acid sequence studies can still produce significant findings. Comparisons between homologous genes can be classified in 3 ways: Orthologous comparison results from equivalent genes in different taxonomic groups, Paralogous from equivalent genes in the same organism, and Metalogous from non equivalent genes in different taxonomic groups (e.g. α-and β-globins between mouse and man) (Ferguson, 1980).

Alignment of the amino acid sequences of orthologous polypeptides enables the determination of a difference measure at this level, which is equivalent to a measure of genetic distance since the two sequences shared a common ancestor. If multiple pairwise comparisons are made with a group of orthologous polypeptides, phenetic (overall similarities at the present time (Sneath and Sokal, 1973)) or cladistic (sequential ordering of splitting of individual lineages, (Henning, 1966)) relationships can be deduced. After constructing a matrix of values resulting from each comparison, phylogenetic trees can be drawn up from such data, using a variety of sophisticated algorithms, which generally seek to best represent the information, while at the same time achieving maximum parsimony (i.e. seeking to minimise difference), in ways that best reflect the way the sequences are related and have descended from one another (Ferguson, 1980). A variety of methods have been used to produce such trees from globins (Czelusniak et al., 1982), fibrinopeptides and carbonic anhydrases (Goodman, 1973, 1974) and cytochrome c's (Peacock and Boulter, 1975) across a wide range of taxonomic groups.

1.2 Seed Storage Proteins

1.2.1 Major features of seed storage proteins

Proteins of seeds fall into two major catergories, namely the storage proteins, which are the major components of total seed protein and the so-called 'housekeeping' proteins essential for normal functioning of cell metabolism. A system of classification has been defined to classify proteins from seeds and grains on the basis of their solubility (Osborne, 1924):

(i) Those soluble in water - the Albumin class.

(ii) Those soluble in dilute salt solutions at neutral pH (phosphate or borate-buffered saline (0.15M NaCl) at pH 7.5 is often used (Croy and Gatehouse, 1985)) - the Globulin class.

(iii) Those soluble in alcohol (50% Propan-1-ol or 55% Propan-2-ol now preferred (Shewry et al., 1984)) - the Prolamin class.

(iv) Those soluble in alkali (an ill-defined term, generally used only for proteins soluble in strongly denaturing solvent (Croy and Gatehouse 1985) - the Glutelin class.

To adopt the components produced by such solubility schemes under the blanket term of seed storage protein would not be wholly correct. The seed storage proteins must also conform to several definitions (Pernollet and Mossé, 1983); Generally, these proteins must supply the source of reduced nitrogen needed during germination and early seedling growth. Consequently such components must be observed to be rapidly degraded at the onset of germination (Boulter, 1982). They accumulate in a tissue-specific manner in endosperm or cotyledon tissue during seed development. They are found in relative abundance to other seed nitrogen compounds - indeed they may constitute as much as 70% of seed dry weight (Croy and Gatehouse, 1985). They are localised in storage organelles called protein bodies, surrounded by a single membrane of tonoplast or endoplasmic reticular origin (Pernollet, 1978, Weber and Neumann, 1980) this packaging affording protection during seed development. Finally, they usually show, especially in the case of legume and cereal seed storage proteins, unusual profiles of amino acid composition e.g. in legumes they are often rich in the amino acids containing more nitrogen atoms - Aspargine, Glutamine, Arginine and Proline, with a consequent lack of amino acids such as Cysteine, Methionine and Tryptophan (Higgins, 1984).

Generally the major storage proteins of legumes and many other dicots are the globulins, whereas in monocots. they are prolamins and glutelins (oats being the only exception, having a high globulin content (Higgins, 1984)). The storage globulins are usually present in the seed as oligomeric molecules, and separation of constituent polypeptides after denaturation on SDS-PAGE shows a large degree of heterogeneity in terms of size, charge and composition. This complexity is consistent with the large numbers of genes thought to code for these proteins (see section 1.2). An example is pea legumin or vicilin, shown to have up to 30 separable polypeptides under two-dimensional IEF/SDS-PAGE (Matta et al., 1981)

At this point, considering the major difference between the seed storage genes of cereals and legumes, those of cereals will not be discussed further here. (Reviews on this topic-see Miflin and Shewry, 1979, Miflin and Shewry, 1981, or Payne and Rhodes, 1982.)

1.2.2 Seed Storage Protein and Gene Structure of some Legumes

1.2.2.1. Pea, *Pisum sativum* L.

Pea storage proteins are comprised of two main immunologically distinct classes, the 11S, legumin and 7S, vicilin globulins. Together, these classes contribute up to 70% of seed dry weight (Croy and Gatehouse, 1985). These proteins are synthesized on membrane bound polysomes, cytologically known as the rough endoplasmic reticulum or R.E.R. (Bollini and Chrispeels, 1978). They are then packaged into the protein bodies.

i) Legumin

Legumin is a hexameric molecule of Mr 360,000 - 400,000, consisting of 6 sub-unit pairs, each of which consists of a disulphide-linked acidic (α) (Mr - 38,000) and basic (β) (Mr - 21,000) polypeptide pair. Evidence from one and two dimensional gel electrophoresis has shown the α -polypeptide to exhibit considerable heterogeneity in both size and charge, while the β -polypeptide only shows major differences in charge, its size showing only minor variations (Casey, 1979a, 1979b, Croy et al., 1979, Krishna et al., 1979). Based on this, a classification has been developed (Matta et al., 1981) to describe minor legumin classes with larger α polypeptides (big legumin, Mr 41,000 - 42,000), each associated with specific β polypeptides, as opposed to the major legumin polypeptides with α subunits of Mr - 38,000.

Synthesis of the α and β subunits is known to occur in the form of a 60,000 Mr preproprotein precusor, containing both the α and β polypeptide chains and also a 21 amino acid leader sequence (Evans et al., 1979, Croy et al., 1980, Lycett et al., 1984). The molecule is then subjected to post-translational proteolytic cleavage and the resulting mature polypeptide chains then associate via non-covalent bonding to form the mature legumin molecule. Such a model of precursor-product relationship has now been extended and confirmed to operate in the synthesis of all legumin-type storage proteins, mainly from the results of pulse-chase experiments (Chrispeels et al., 1982a, 1982b).

Nucleotide sequence data from legumin cDNAs revealed several other significant features (Croy et al., 1982, Lycett et al., 1984b). Firstly, it was observed that the α polypeptide is synthesized before the β ,as it is encoded in the 5' region of the messenger RNA. Secondly, the presence of 3 tandem repeats of 18 polar amino acid residues was revealed in the C terminal region of some, but not all cDNAs. Lastly, the implied amino acid sequence from the data suggested a potential cleavage site for the α and β polypeptide chains adjacent to the N-terminal amino acid of β , where the sequence Asn-Gly conforms to the cleavage site found in several other storage proteins which have an exposed Asn-X bond (Lycett et al, 1984). Indeed a hydrophilicity profile of legumin (a prediction of secondary structure based on varying degrees of hydrophilicity in the molecule) shows the Asn-Gly bond possibly lying on the surface of the molecule, where it would be accessible to cleavage (Croy and Gatehouse, 1985).

A final important feature of pea legumin is that it is non-glycosylated (Croy et al., 1979), consistent with legumin proteins in general, which are known to contain little or no carbohydrate (Derbyshire et al., 1976).

Recent studies on the legumin gene family involving hybridisation with two new legumin cDNAs, pCD32 and pCD40 (Domoney and Casey 1984), combined with earlier results using a different cDNA pRC 2.2.4., (Croy et al., 1982) indicate the possible presence of 8 legumin genes in the haploid pea genome. At least 4 of these genes appear to code for the major 60,000 Mr polypeptides (Croy et al., 1982), with the others apparently representing 63,000, 65,000 and 80,000 Mr minor legumin species (Domoney and Casey, 1984). It would be expected that a copy number greater than eight would be the only way to account for the existence of 'small' legumin polypeptides with α subunits of 24-25,000 Mr (see above).

Nucleotide sequence data from Legumin gene A (Lycett et al., 1984) shows the presence of all the consensus sequences characteristic of eukaryotic genes. In relation to the transcription start (34 bp upstream from the initiation codon), a TATA box at -30, a CAAT box at -87, and a potential AGGA box at -103. Additionally the data enabled identification of regions homologous to viral enhancer elements (see section 1.1.2.2.). At the 3' end of the gene are 3 sets of double overlapping polyadenylation signals. All polyadenylated cDNAs so far examined have a poly (A) tail 19-20 bp downstream from the second of these signals (Lycett et al., 1984), with a similar situation also having been observed in a soybean lectin gene (Vodkin et al., 1983).

The sequence data also reveals the presence of 3 short introns, two in the region coding for the α polypeptide, both of 88 bp, and one in the β polypeptide coding region, of 99 bp. All 3 introns obey the GT/AG boundary rule (Breathnach et al., 1978).

Chromosome mapping studies (Matta and Gatehouse, 1982, Domoney et al., 1986) refer to chromosomal locations of the legumin genes, one being the short arm of chromosome 7, near the r locus, termed Leg-1, and another (minor) one on chromosome 1, near the a locus, termed Leg-2.

(ii) Vicilin

Unlike the 11S legumin fraction of pea, the 7S vicilin fraction is not essentially homogeneous in its composition, and evidence (Derbyshire et al., 1976) suggested it contains more than one major protein. This was confirmed when a third storage protein, convicilin, immunologically related to vicilin, was purified from the 7S fraction using non-dissociating techniques (Croy et al., 1980).

The mature vicilin molecule has an Mr value of 145,000 - 170,000, resulting from the trimerisation of 3 50,000 Mr vicilin polypeptide subunits (Gatehouse et al., 1981). However, SDS-PAGE analysis shows the presence of a complex array of constituent polypeptides ranging from Mr 50,000 to species of 33,000, 19,000, 16,000, 13,500 and 12,500 Mr, along with minor species of 35,000 and 31,000 Mr (Croy et al., 1980 b., Chrispeels et al., 1982b).

Pulse-chase studies on the assembly of vicilin protein oligomers (Chrispeels, et al., 1982b), shows the 50,000 Mr species to carry the majority of the radioactive label. During the chase period, the label begins to appear in the smaller vicilin polypeptides, and has also now moved from the R.E.R to the protein bodies. A simultaneous disappearance of one of the 50,000 Mr subunits, (47,000) clearly suggested a precursor-product relationship. Thus an overall model of trimerisation of 50,000 Mr subunits some of which are then susceptible to post translational cleavage or 'nicking' whilst still maintaining overall structural integrity was built up (Gatehouse et al., 1981). The presence of full size 50,000 Mr subunits in the mature protein was explained by the lack of sites susceptible to cleavage in such polypeptides. Later studies on vicilin cDNAs (see below) confirmed that two potential cleavage sites were present in some vicilin 50,000 Mr precursors (reviewed by Boulter, 1984).

Unlike legumin, pea vicilin was found to be glycosylated (Chrispeels et al., 1982a), the process occurring in the lumen of the R.E.R. Carbohydrate has been found in the 50,000, 26,000 and 14,000 Mr polypeptides (Badenoch-Jones et al., 1981), and has been attributed as causing the difference in size between the 14000 and 12500 Mr polypeptides (Davey et al., 1981).

Analysis of vicilin cDNA nucleotide sequence (Gatehouse et al., 1982b, 1983, Lycett et al.,1983a) plus its comparison with polypeptide sequence data revealed other important features. Firstly, from the sequence of a nearly full length 47,000 Mr cDNA, the presence of a 15 amino acid residue leader sequence was confirmed, which facilitates transport of the nascent polypeptide into the lumen of the R.E.R. (Blobel and Dobberstein, 1975). Secondly, comparison of predicted amino acid sequences with those produced by direct polypeptide sequencing implied the determination of potential cleavage sites by amino acid sequence specificity. Two sites, α : β and β : γ have been identified, and it was suggested that the amino acid sequence Gly-Lys-Glu-Asn immediately prior to the site allows proteolysis, with the presence of Asn apparently crucial (Gatehouse et al., 1982b, 1983). Serological studies plus comparison of N terminal and cDNA predicted amino acid sequences has confirmed this theory (Spencer et al., 1984). A hydrophilicity profile produced from amino acid sequence predicted from a vicilin cDNA implies these sites lie on polar, surface regions of the molecule (Croy and Gatehouse, 1985).

These cDNAs contained no 5' flanking sequence and only one (coding for 50,000 Mr vicilin) contained any 3' flanking sequence and showed the presence of only 1 polyadenylation signal.

At the outset of this project less was known about the vicilin gene family than that of legumin. It was thought to consist of at least 11 genes (Domoney and Casey, 1984), located on 6 different loci in the genome, 3 of these loci, vc-1, vc-2, vc-3, located on chromosome 7 near the Leg-1 locus (Ellis et al., 1986). The genes code for both 50,000 and 47,000 Mr polypeptides.

Even less is known about convicilin and the gene(s) that code for it. Allied to the fact that convicilin is often present in preparations of vicilin, is the likelihood of 'hybrid' molecules composed of vicilin and convicilin polypeptides (Gatehouse et al., 1981).

The mature convicilin protein has an Mr of 210,000-280,000 consisting of 3 or 4 70,000 Mr subunits (Croy et al., 1980) which show little heterogeneity and are not disulphide bonded (Croy et al., 1980b). A cDNA for convicilin has been isolated (Domoney and Casey 1983) and used to show that the mRNA coding for convicilin is in a different (larger) size class to that of vicilin. The chromosomal location of the convicilin gene(s) has been mapped to chromosome 2 close to the k locus (Matta and Gatehouse, 1982).

1.2.2.2 Soybean

The major storage proteins of soybean *Glycine max* are the globulins - glycinin (12.2S) and β conglycinin(7S). The ratio of glycinin to β -conglycinin varies according to cultivar, from 3:1 to 1:1 (Nielsen, 1984, Pernollet and Mosse, 1983). Together they constitute approximately 70% of total seed protein. Glycinin and β -conglycinin can be considered as homologues of legumin and vicilin of pea respectively, and as such show a similar passage through the cell, with synthesis on membrane bound polysomes and eventual storage in cotyledon protein bodies (Barton et al., 1982, Sengupta et al., 1981).

(i) Glycinin

Mature glycinin has an Mr value of 360,000, and is a hexamer of 60,000 Mr subunits. Each subunit consists of an acidic (A or α) polypeptide, Mr 37,000, disulphide bonded to a basic (β or B)polypeptide, Mr 22,000 (Catsimpoolas et al., 1971). There are 5 major different subunit groups, showing compositional polymorphism (Kitamura et al., 1980). The polypeptides show extensive size and charge heterogeneity, with at least 6A and 5B having been separated and purified (Moreira et al., 1979). Consistent with the system of precursor-product found in other legumins (Bassuner et al., 1983, Brinegar and Peterson, 1982, Croy et al., 1980), the pairing of A and B subunit polypeptides has been shown to be non-random, implying a precursor containing both polypeptides is initially produced and is then subject to post translational cleavage in the protein bodies (Staswick et al., 1981, Turner et al., 1981).

Strong evidence for the presence of a leader sequence has been obtained by comparing the size of glycinin subunits obtained by *in vitro* translation of messenger RNA in both wheat germ translation systems (Barton et al., 1982) and rabbit reticulocytes (Turner et al., 1981), with those subunits produced *in vitro* , which proved to be smaller. It seems likely that this is due to a post translational cleavage of signal peptide *in vivo* . Subsequent analysis of glycinin cDNAs suggest leaders of 19-24 amino acids (Negoro et al., 1985, Fukazawa et al., 1985). This data also showed the presence of a 4 amino acid linker between the A and B polypeptides at Lys_{278} and Arg_{279}, suggesting a similar cleavage site to that of the C peptide in proinsulin (Docherty et al., 1982).

Comparison of the amino acid sequences (predicted from cDNA sequence data and also directly obtained by protein sequencing) shows close homology between the N terminal sequences of the glycinin B polypeptide and the corresponding *Vicia faba* legumin subunit polypeptide (Gilroy et al., 1979) and also the N-terminal sequences of glycinin A and pea legumin α (Casey et al.,1981).

A final interesting point has been noted that regions of internal homology within the glycinin α subunit might reflect an analogous situation to the repetitive block structure of the cereal prolamins (Moreira et al., 1981).

Reports of the number of genes in the glycinin gene family suggest a figure of 4 (Moreira et al., 1981) or possibly 5 (Fischer and Goldberg, 1982). As with pea legumin, the genes appear to encode both A and B polypeptides with A being produced from the 5' end of the gene (Neilsen, 1984).

Nucleotide sequence from a glycinin genomic clone corresponding to the primary structure of the A_2 B_{1a} polypeptide (Thanh et al., 1984), together with S1 mapping data, show the transcription start at 43 bases upstream from the ATG initiator codon, with a TATA box at -25, and a CAAT box at -115. At the 3' end of the gene there appear to be 3 potential polyadenylation signals, with the poly(A) tail in A_2B_{1a} message starting 15-17 bases after the last of these signals. The gene also contains 3 introns of 238, 292 and 624 bp respectively. Although these are larger than those of legumin in pea, their positions correspond (Lycett et al., 1984). The intron/exon boundaries conform to the GT/AG rule (Breathnach et al., 1978).

Little is known of the chromosomal location of these genes, apart from some evidence that they do not appear to be closely linked to each other, and that, suprisingly some may be linked to leaf tissue - specific genes (Fischer and Goldberg, 1982).

ii β conglycinin

In soybean there are three immunologically distinct conglycinins; α conglycinin-a 2S monomer, γ conglycinin - a minor component of the 7S globulin fraction and β conglycinin (also known as soybean vicilin) - the major 7S globulin (Catsimpoolas and Ekenstam, 1969).

β, conglycinin, like pea vicilin, is glycosylated at an average level of 5% (Thanh and Shibasaki, 1977), and, although antigenically unrelated, it bears a close resemblance to vicilin proteins of other legumes (Derbyshire et al., 1976).

It has an Mr value ranging from 140,000 to 210,000 and is a trimeric molecule composed of 3 subunits; α -Mr 54,000-76,000, α′-Mr 54,000-86,000 and β -Mr 40,500-53,000. All 3 are acidic and have very similar amino acid composition (indeed all are devoid of cysteine). The β conglycinin subunits appear to undergo a very complex process of co- and post-translational modification. Leader sequences are cleaved cotranslationally by the endoplasmic reticulum (Sengupta et al., 1981). The β subunit is synthesised later in development than α and α ' (Gayler and Sykes, 1981) and all 3 undergo complicated processes of glycosylation and cleavage which will not be considered further here.

cDNAs from genes coding for the α and α ' subunits have been isolated (Schuler et al., 1982ab) but as yet, none corresponding to the β subunit have been found. In all, 19 cDNAs were isolated, and on the basis of their nucleotide sequence were clearly segregated into two groups representing the α and α ' polypeptides. The two classes of sequences differed by only 6% in their nucleotide sequence, and as such represent two closely related multigene families (Schuler et al., 1982). This report also presented partial DNA sequence from a genomic clone representing an α ' subunit gene. No 5' sequence was present, but the data did show the presence of 4 small introns (as opposed to 1-2 predicted by R-loop mapping, Fischer and Goldberg, 1982) of 85, 115, 132 and 40 bp, whose borders conform to the intron/exon boundary rule (Breathnach et al., 1978). Also the presence of only 1 double overlapping polyadenylation signal was revealed, which appears to lie in region of potential secondary structure, with the poly(A) tail attached 29-32 bases downstream from here in most of the cDNAs (one exception being a poly(A) tail only 17 bases downstream in one α subunit cDNA).

Gene copy number studies are complicated by the subunits of mature β conglycinin being encoded by separate multigene families. An overall estimate of 5 copies of 7S subunit genes has been made (Goldberg et al., 1981) using probes common to all 3 subunits. Isolation of 3 different genomic clones (Goldberg et al., 1983a) each containing 2 separate β conglycinin genes demonstrates the clustering of the genes in these multigene families, but no further evidence on chromosomal location or arrangement has been reported other than that the genes are present on at least 3 different loci (Davies, 1985).

1.2.2.3 Frenchbean

French bean (*Phaseolus vulgaris*), unlike pea and soybean, has only one major seed storage protein, phaseolin, a 7S globulin. It accounts for up to 50% of the protein in mature seeds (Ma and Bliss, 1978) and has an Mr value of 140,000-160,000 Mr (Derbyshire et al., 1976). One dimensional SDS-PAGE reveals the presence of 3 distinct polypeptide subunits - α (51-53,000 Mr), β (47-48,000 Mr) and γ (43-46,000 Mr) (Hall et al., 1977, Brown et al., 1981a, 1981b), and evidence has been presented (Pusztai and Stewart, 1980), that these subunits are arranged in trimeric form, which, along with 3 - 5% glycosylation (Hall et al., 1978) constitute the phaseolin molecule.

Phaseolin subunits are synthesized on membrane bound polysomes (Bollini and Chrispeels, 1978) and have leader sequences which are cotranslationally removed by the E.R. (Bollini et al., 1983) allowing the polypeptides to be sequestered into the lumen (Baumgartner et al., 1980). A system of glycosylation and proteolytic processing analogous to that occurring in β -conglycinin (Sengupta et al., 1981) then appears to occur (Bollini et al., 1983).

Two dimensional electrophoresis of phaseolin resolves 5 polypeptides, indicating both size and charge heterogeneity in the phaseolin protein pool (Brown et al., 1981a). However, peptide mapping of these phaseolin proteins after proteolytic and chemical cleavage implies that they are all highly homologous (Ma et al., 1980, Bollini and Vitale, 1981), strongly suggesting that they are encoded by a multigene family. Since then, a figure of 7 genes per haploid genome has been reported (Talbot et al., 1984). In the developing cotyledon, phaseolin polypeptides have been found to be encoded in a 16S mRNA species representing approx. 40% of total poly(A) RNA (Hall et al., 1980, Murray and Kennard, 1984). cDNA clones complementary to this messenger RNA have been isolated and their nucleotide sequence determined (Slightom et al., 1985), demonstrating that the phaseolin polypeptides are encoded by two distinct phaseolin gene subfamilies, termed α and β phaseolin, with α phaseolin polypeptides encoded by genes containing direct repeats coding for an additional 14 amino acids. Apart from this, the two gene subfamilies show 98% homology.

Sequence analysis of a β -type phaseolin genomic clone (Hall et al., 1983) indicates the presence of 5 small introns, which all conform to the intron/exon boundary rule (Breathnach et al., 1978). Interestingly this data also reveals the presence of 3 possible TATA boxes and 2 CAAT boxes (possibly analogous to a double promotor system found in zein proteins of maize (Langridge and Feix, 1983)), and these are believed to be responsible for the numerous cap sites found in phaseolin mRNAs (Slightom et al., 1985)

Genetic analysis of *P.vulgaris* cultivars which exhibit different two-dimensional PAGE protein patterns show no recombinant protein phenotypes, suggesting a close linking of the genes (Brown et al., 1981b), however at the molecular level, distance between these genes could be considerable (Sun et al., 1981, Talbot et al., 1984).

1.2.2.4 Broad bean

A brief mention will be made here of the storage proteins of the broad bean (*Vicia faba*), since although reference is made later in the text, little documented information has been published on this subject.

Vicia faba var. minor possesses two sub-families of 11S legumin genes, termed A and B. The major subunits of the *V.faba* legumin, A and B, were initially distinguished according to the presence (A-type) or absence (B-type) of Methionine, and subsequent to this, amino acid analysis, sequence determination and peptide mapping (Horstmann, 1983). Finally at the DNA level, hybridisation studies and especially cDNA sequence data confirmed that *V.faba* did indeed possess a legumin multigene family, subdivided into A and B subfamilies (Wobus et al., 1986).

Such an occurence of two gene subfamilies for legume 11S storage proteins is also found in soybean (Neilsen, 1984, Scallon et al., 1985) and pea (Wobus et al., 1986).

A genomic clone, representing a B-type *V.faba* legumin has been isolated (Baumlein et al., 1986), and nucleotide sequence data from this shows in the 5' flanking sequence, a cap site which appears conserved in location and sequence between several other plant genes (Vodkin et al., 1983, Messing et al., 1983). 30 bp upstream from here is a TATA box. At the 3' end of the gene are 3 putative polyadenylation signals, with a poly(A) tail 23 nucleotides downstream from the 3rd of these elements found in a cDNA clone pVfc 70.

In the coding sequence, this data gave evidence of a 21 amino acid residue signal peptide, and the presence of two short introns of 95 and 100 bp. These were designated introns 2 and 3, as they exactly occupy the positions of introns 2 and 3 in the A-type legumin genes of soybean and pea (see sections 1.2.2.1, and 1.2.2.2) Each intron obeys the normal intron/exon boundary rule (Breathnachet al., 1978).

1.3 Aims and Objectives of the Project

In order to further the knowledge of pea seed storage protein gene structure and regulation, selected genomic clones coding for legumin and vicilin were fully characterised, and their nucleotide sequence determined. Subsequent to this, comparisons of the sequence of 11S and 7S genes of main legume species were hoped to show areas of sequence important to gene regulation, integrity and tertiary structure.

Additionally, it was intended that nucleotide sequence data determined in the project combined with any such data already determined on legume seed storage proteins, might be used to establish, in the most appropriate way, the evolutionary pathway in which the current multigene families for such proteins, have developed.

Finally, as a partial augmentation of these potential evolutionary findings, it was hoped by using material from ancient or primitive pea lines, the pattern of development of multigene families would be elucidated, along with any relationship between storage protein gene copy number and the levels of storage protein deposition in legume seed cotyledons.

Sl

CHAPTER 2 : MATERIALS AND METHODS

2.1 Materials

2.1.1 Biological and Chemical Reagents

All reagents, with the exception of those noted below, were obtained from BDH Chemicals Ltd., Poole, Doreset, U.K., and were either of analytical grade or the best available.

Acridine Orange, Ampicillin, Adenosine triphosphate (ATP), Bovine Serum Albumen (BSA), Dithiothreitol (DTT), Ethidium Bromide (Et Br) Glyoxal, Herring Sperm DNA, I PTG, Lauroyl Sarcosine, Lysozyme, N,N'-Methylene-bis-Acrylamide, Polyvinyl pyrrolidone (PVP), Pronase P, RNAase A, Spermidine, and t-RNA were from Sigma Chemical Co., Poole, Dorset, U.K.

Boric Acid, Caesium Chloride, Sodium Chloride and Sodium Dihydrogen Orthophosphate were from Koch-Light Ltd., Haverhill, Suffolk, U.K.

Ficoll 400, Klenow fragment, M13 17-mer primer, Sephadex G-50 and T_4 DNA Ligase were from Pharmacia Fine Chemicals, Uppsala, Sweden.

Nitrocellulose filters (BA 85, 0.45 um) were from Schleicher and Schull, Anderman and Co. Ltd., Kingston-upon-Thames, Surrey, UK.

Whatman 3MM paper and 2.5cm GFC Discs were from Whatman Ltd., Maidstone, Kent, UK.

Bacto-Agar, Bacto-tryptone and Bacto-Yeast Extract were from Difco Laboratories, Detroit, Michigan, USA.

BBL-Trypticase Peptone was from Becton Dickinson and Co., Cockeysville, Maryland, USA.

Restriction enzymes were obtained from Bethesda Research

Laboratories (BRL), Bethesda, Maryland, USA, Northumbria Biochemicals Ltd., (NBL), Cramlington, Northumberland, UK, Pharmacia Fine Chemicals and the Boehringer Corporation (London) Ltd., Lewes, East Sussex, UK.

Alkaline Phosphatase, Polynucleotide Kinase, Tris (hydroxymethyl) aminomethane (Tris), and Xgal were from The Boehringer Corporation (London) Ltd.

Agarose and Low Melting Point (LMP) Agarose were from BRL.

DNA Sequencing Kits (N 4502), Nick translation kits (N 5000) and Radionucleotides were from Amersham International plc., Amersham, Bucks, UK.

Gel bond was from ICN Biomedicals Ltd., High Wycombe, Bucks, UK.

Leaf Genomic DNA from *Pisum sativum* L var. Feltham First and Poly A+ Cotyledon RNA were gifts from Dr. J.A. Gatehouse and D. Bown.

Plasmids and λ NM258 were supplied by Dr. R. Croy from communal departmental stocks.

All solutions (except those containing SDS, and those used as electrophoresis or Southern and Northern blotting buffers) were sterilised by autoclaving.

2.1.2 Growth of Biological Material

Seeds were germinated on damp tissue paper for 3-4 days, in darkness, at room temperature. After this time they were planted in 15cm pots filled with Levington Potting Compost. Watering was at 3 day intervals, and the plants grown to maturity under greenhouse conditions. Leaf material was harvested and stored in silver foil at -80°C.

2.1.3 Growth Media for Bacteria and Bacteriophage

Table 1 presents a list of the nutrient composition of the various media used.

TABLE 1 COMPOSITION OF GROWTH MEDIA

Medium	Nutrient Composition (per litre)
YT Broth	8g Trypticase-peptone 5g Yeast extract 5g Na Cl
2xYT Broth	16g Trypticase-peptone 10g Yeast extract 5g Na Cl
YT Agar	As YT Broth, plus 25g Bacto-agar
YT Top layer Agar	As YT Broth, plus 7.5g Bacto-agar
YT/Amp/Xgal	As YT Agar, plus 50 ug/ml Ampicillin and 40 ug/ml Xgal

2.1.4 Bacterial Strains and plasmid and bacteriophage vectors

Bacterial strains were derivatives of *E.coli* K12. Table 2 lists these strains, plus plasmid and phage vectors, with sources or references accompanying each one.

TABLE 2 E. Coli Strains, Plasmids and Bacteriophage

Bacterial strains	Genotype	Ref. or source
JM 83	ara, Δ (lac-proA,B,), rpsL(=strA) ϕ 80,lacZ Δ M15	Bethesda Research Laboratories (B.R.L)
JM 101	supE,thi, Δ (lac proA,B) /F' traD36,proA,B,laql^{q}Z Δ M15	Yanisch-Peron et al, 1985 or B.R.L.
Plasmids		
pDUB 2	Ap^{R} Vic 50K(in pBR322)	Croy et.al, 1982 Lycett et.al, 1983a.
pDUB9	Vic 50K " "	Delauney 1984
pDUB21	Leg A (in pUC 8)	Lycett et al 1984
pDUB24	Leg A " " "	Lycett et al 1984
pUC8	Ap^{R} lacZ	Vieira and Messing 1982
pBR322	Ap^{R} Tc^{R}	Bolivar et al 1977
Bacteriphages		
M13 mp8	Multiple cloning site	Messing 1983
M13 mp9	" " "	" "
M13 mp18	" " "	Yanisch-Peron et al 1985
M13 mp19	" " "	" " " "

Key: Ap^{R} = ampicillin resistance, Tc^{R} = tetracycline resistance
vic 50K = 50K vicilin cDNA
LegA = Legumin Gene A

N.B. For genomic clones used see Results, Section 3

Methods Section

2.2 Glass and Plasticware

All glass and plasticware was thoroughly washed in Teepol solution, rinsed once in tap water and a further 3 times in distilled water, before use. Any vessels coming into contact with DNA or RNA were siliconised with Repelcote (2% Dimethyldichlorosilane in 1.1.1. Trichloroethane) rinsed once with distilled water, and then autoclaved. All Eppendorf tubes were also siliconised and autoclaved prior to use.

2.3 Phenol Extraction and Ethanol Precipitation of DNA

Extraction of protein (including modifying enzymes), removal of metal ions and precipitation of DNA in large volumes of solution was achieved by Phenol Extraction and Ethanol precipitation.

2.3.1 Phenol Extraction

Firstly, either water or TE buffer (10mm Tris, 1mm EDTA, adjusted to pH8.0 with HCl) was added to ensure a minimum sample volume of 200ul. An equal volume of Phenol (saturated with TE buffer) + 0.1% Hydroxyquinoline (preventing the oxidation of Phenol in 4°C storage) was added, the sample vortex mixed and then spun at 12,000g. The aqueous phase was then re-extracted in the same way in a fresh tube. The original phenolic phase was then mixed with an equal volume of TE buffer and after vortexing and centrifugation, the two resulting aqueous phases combined in a fresh tube. An equal volume of Choroform/Isoamylalcohol (24:1) was then added to this combined aqueous phase followed by vortexing, and a brief centrifugation step to separate phases. This step was then repeated to ensure the complete removal of any phenol. The final aqueous phase could now be precipitated with Ethanol.

2.3.2 Ethanol Precipitation

A 1/10th volume of 3M Sodium Acetate, pH5.2 (with Acetic Acid) was added to the sample and mixed, followed by the addition of 2 volumes of -20°C 100% Ethanol for a minimum of 1 hour. The DNA was then pelleted by spinning at 12,000g for 10 mins, the ethanol discarded, and the DNA sample washed with -20°C 80% Ethanol (each wash being followed by spinning for 7 mins at 12,000 g after vortex mixing). Finally, the Ethanol was drained off, and the sample dried under vacuum for 10 minutes (with perforated Nescofilm stretched over the sample vessel). The resulting DNA pellet was then ready for resuspension.

2.4 Use of DNA Modifying Enzymes

2.4.1 Restriction Endonucleases

2.4.1.1 Plasmid DNA

The unit definition for any restriction endonuclease enzyme is that 1 unit will digest 1 ug of DNA to completion in 1 hour. Generally, a 2-3 times unit excess was used over a 2 hour incubation period to digest DNA of plasmid origin. Digests were always performed in a volume at least 10 times that of the enzyme volume used (this ensures adequate dilution of otherwise inhibitory glycerol present in the enzyme solution) and under the buffer conditions recommended for the particular enzyme by its manufacturer. Where necessary RNase (previously boiled for 10 minutes to destroy nucleases) was included in the digest to a final concentration of 1ug/ul.

2.4.1.2 Genomic DNA

For Genomic DNA, an enzyme excess of at least 5 times was used, over a digestion period of 3 hours, with the addition of nuclease free BSA to a final concentration of 0.1ug/ul.

2.4.2 T_4 DNA Ligase

Ligation of DNA fragments was performed in buffer concentrations of 0.6mM ATP, 0.066M Tris/HCI pH7.5, 0.01mM $MgCl_2$, 0.01mM β -Mercaptoethanol with 1 unit of T_4 DNA ligase (Weiss et al., 1968), usually in a final reaction volume of 10 ul. Incubation was for either 2 hours at room temperature, or overnight at 15°C. Such conditions proved suitable for ligation of both 'sticky' and 'blunt' ended DNA fragments.

2.4.3 Alkaline Phosphatase

5'-dephosphorylation using Calf Intestinal Phosphatase was performed in buffer conditions of 50 mM Tris/HC1 pH 9.0, 1mM $MgC1_2$,0.1 mM $ZnC1_2$, 1mM Spermidine. 5'-dephosphorylation of 5' overhanging ends was carried out over 1 hour at 37°C, using an enzyme concentration of 0.2 units/ug. For blunt ended molecules, an incubation of 15 minutes at 37° was followed by one at 56°C for the same period, after which a second aliquot of enzyme was then added and these incubation steps repeated.

2.5 Agarose Gel Electrophoresis

2.5.1 Agarose Gels

Agarose gels of varying percentages (see Maniatis et al., 1982 for details) were used to separate, identify and isolate DNA fragments. Agarose was dissolved by boiling in 200 mls of Tris/Acetate EDTA buffer (0.04 M Tris/Acetate, 0.001M EDTA, pH7.7 with glacial acetic acid). Et Br was added to a final concentration of 1ug/ml, and the gel cast in a 150 x 185 mm perspex surround with suspended well former attached to a polished glass plate with silicone grease. Once set, the surround and well former was removed, and the gel placed in a horizontal submarine electrophorisis tank, which was then flooded with 2.1 litres of Tris/Acetate EDTA buffer plus 1ug/ml Et Br. Samples were mixed with ½ vol. Agarose beads (31.25% Glycerol, 10mM Tris/Mcl pH8.0, 0.01M EDTA pH8.0, 0.02% Agarose, plus 0.01% Bromophenol Blue or Fast Orange G - Autoclaved and then extruded through a fine needle several times to prevent setting) and loaded into the wells. Electrophoresis, in the direction of the positive terminus, was either for 4 hours at 100-120 V, or overnight at 35V.

2.5.2 Minigels

50 mls of gel solution (0.5-0.8% Agarose, TBE buffer - 0.089M Tris, 0.089M Boric Acid, 0.025M EDTA, plus 4ug/ml EtBr) was cast in a Minigel Apparatus (Cambridge Uniscience). 50 mls of TBE buffer plus 4ug/ml Et Br was then added, samples loaded as in section 2.5.1, and electrophoresis performed at 60 mA for 1 hour.

2.5.3 LMP Agarose gels

LMP-Low Melting Point - Agarose gels were cast in a 50 x 185 mm perspex surround. Usually 60 mls of gel solution (0.8% LMP Agarose, 0.04 M Tris acetate, 0.001M EDTA, pH7.7 plus 1ug/ml Et Br) was used. Sample loading and electrophoresis were as detailed in section 2.5.1.

2.5.4 Visualisation and Photography of DNA in Gels

Once electrophoresis was completed, gels were transferred onto a U.V. transilluminator, λ 254 nm, for DNA visualisation. Where necessary, photographs were taken with a Polaroid Land camera, using 3,000 ASA Type 667 Polaroid film, at f16 for 5 seconds, through an orange filter.

2.5.5 Glyoxal RNA Gels

The method used was that of McMaster, 1977. Firstly, RNA samples were glyoxylated by combining, in the following order, in a 0.75 ml Eppendorf tube, 25 ul of redistilled Dimethyl Sulphoxide (DMSO), 2.5 ul of 0.2M Sodium Phosphate buffer (pH 7.0), 30% glyoxal (deionized with Amberlite resin for 1 hour under Nitrogen) and finally the RNA sample in 15.3 ul water. Incubation followed for 1 hour at 50°C.

Next, a 1.5% High Gelling Temperature (HGT) Agarose gel was prepared in 120 mls of 10 mM Sodium Phosphate Buffer, pH 6.8. The gel was cast on the hydrophilic side of a sheet of Gel bond (FMC corporation) cut to fit inside an electrophoresis tank. An adjustable well former was used to ensure deep wells.

To run the gel, a submarine electrophoresis tank with no legs was placed on two magnetic stirrers on a levelling table. The gel was placed in the tank, two glass rods placed along its sides to prevent curling of the Gelbond, and the tank flooded with 10 mM Sodium Phosphate Buffer, pH 6.8. Agarose beads (Section 2.5.1) made up in running buffer were added to the samples before loading. A peristaltic pump was then set up to circulate the buffer during electrophoresis, which was carried out at 120V for 3 hours with slow stirring.

Once run, the gel was stained in freshly prepared Acridine Orange (30 mg per litre of running buffer) for 15 minutes in the dark at 4°C, and then destained for 1 hour in running buffer with gentle shaking. Visualisation and photography of RNA was then as in section 2.5.4.

2.6 Isolation of DNA from Agarose Gel

The following methods were routinely used and proved successful.

2.6.1 Glass Fibre Disc (Chen & Thomas, 1980)

The band of DNA was excised from the gel, placed in a preweighed Corex tube, and reweighed. 2 volumes of 8M Sodium Perchlorate were added, followed by vortex mixing until the gel slice had dissolved.

A 6mm Whatman GFC Glass Fibre disc was then placed on 2 layers of Whatman 3MM paper, 4 layers of absorbent blue paper towel and 1 nappy liner. The dissolved sample was then applied to the GFC disc, drop by drop in aliquots of 20ul, ensuring complete absorption of each drop before application of the next. Once all the sample had been applied, the disc was then washed, in exactly the same way with 3 volumes of 6M Sodium Perchlorate followed by 3 volumes of 100% Ethanol. The disc was then air dried, placed in a 0.75 ml Eppendorf tube, thoroughly mixed with 30ul of sterile water, and then incubated for 1 hour at 37° C, allowing the DNA to elute. Finally, a small hole was made in the base of the Eppendorf tube (with a fine gauge syringe needle), and the tube then placed in a large, 1.5 ml Eppendorf tube, and centrifuged for 5 minutes at top speed in a MSE Microcentaur. The resulting DNA solution proved suitable for restriction, ligation and radiolabelling by 'nick translation' (Sections 2.7.1)

2.6.2 LMP Agarose

In this case, the DNA sample was run on LMP Agarose gel, and after excising, weighing and incubating the gel slice at 65°C for 5 minutes, it was then vortex mixed with 2 volumes of extraction buffer (0.5 mM EDTA, 50 mM Tris/HCl pH8.0) and placed at 37°C for 10 minutes. Agarose was then removed by twice extracting with Phenol, and twice with Chloroform/Isoamyl alcohol (24:1) (section 2.3.1.). The DNA was then Ethanol precipitated (section 2.3.2). DNA isolated this way proved less suitable for 'nick translation' (section 2.7.1).

2.6.3 The 'Freeze Squeeze' technique

From the method of Tautz and Renz, 1983, and suitable for use with ordinary agarose gels.

The DNA gel slice was equilibrated with approx 1 ml of 0.3M Sodium Acetate (pH 4.8), 0.01M EDTA for 15 minutes, and then blotted dry on Whatman 3MM paper. A 0.75 ml Eppendorf tube was plugged with sterile siliconised glass wool, the gel slice introduced and the tube placed at -80°C for 20 minutes. Following this, the base of the tube was pierced with a fine gauge needle, and then inside a 1.5 ml Eppendorf, spun for 15 minute at top speed (MSE Microcentaur). The solution thus collected was first mixed with approx 1/20th vol. 1M $MgCl_2$, 10% Acetic Acid, 1/5th vol. 3M Sodium Acetate pH 4.8 and precipitated with 3 vols -20°C 100% Ethanol for 1 hour at -20°C. After a 7 minute spin (again top speed, MSE Microcentaur), the pellet was dried off, Phenol extracted (Section 2.3.1.) and Ethanol precipitated (section 2.3.2). With this method being quick to carry out, give good yields, and provide DNA suitable for all further experiments, it was the one most commonly used.

2.7 α ^{32}PdCTP Labelling of DNA

2.7.1 Nick Translation of Double Stranded DNA

The protocol used was as detailed in the Amersham Nick Translation Kit Booklet (P1/141/84/2) for labelling to a specific activity of 1 x 10^8 c.p.m. per ug of DNA. However, only 0.1 - 0.2 ug DNA were used, and the final reaction volume was reduced to 50 ul. Reagent volumes were reduced accordingly usually with 50 uCi of α ^{32}PdCTP. Incubation was for 2 hours at 15°C, and the reaction stopped by the addition of SDS to a final concentration of 1%.

Separation of labelled DNA from unincorporated α ^{32}PdCTP was achieved on a 20 cm (approx. 7.7 mls vol) column of Sephadex G-50. Approx 1g of Sephadex G-50 was swollen overnight in column buffer (150 mM NaCl, 50mM Tris/HCl pH7.5, 10mM EDTA, 0.1% SDS), and then packed onto the column (plugged at its base with sterile siliconised glass wool). Approx. 0.4 ml fractions were collected in 1.5 ml Eppendorf tubes, and a 1 ul aliquot of each was then taken for scintillation counting (Section 2.7.5) to identify the fractions containing the labelled DNA, and also for calculation of the specific activity of the probe. DNA labelled this way was routinely of a high enough specific activity to probe both genomic and plasmid DNA blots (section 2.8.1)

2.7.2 Labelling Single Stranded Linear DNA By The 'Polymeraid' Method

The Polymeraid labelling kit (P α S biochemicals) was used on very small amounts of single stranded linear DNA (approx 25 ng). Using a α ^{32}PdCTP, specific activities greater than 1 x 10^8 cpm/ug were routinely obtained.

The method used was exactly as outlined on the protocol sheet accompanying the kit.

After running the sample on an LMP Agarose gel, the required band was excised with the minimum amount of extraneous agarose. The gel slice was then placed in a preweighed 0.75 ml Eppendorf tube and sterile H_2O added in the ratio 1.5 ml per gram of agarose. The concentration of DNA was then estimated in ng/ul (the gel slice must be sufficiently small so as this figure does not exceed 1.85 ng/ul, or 25 ng in 13.5 ul - see later), and the sample placed in a boiling water bath for 7 minutes to melt the agarose and denature the DNA. Following this, it was transferred to a 37° water bath for a minimum of 10 minutes. The labelling reaction was then performed in a new 0.75 ml Eppendorf tube by adding 5 ul Polymeraid-RH^{TM} reaction mix, 2 ul Polymeraid-RH^{TM} Klenow fragment, 25 ng (to a maximum of 13.5 ul) of DNA/Agarose mixture, 25 uCi (2.5ul) α ^{32}P-dCTP (@3000 Ci/mmole), 2 ul Nuclease free BSA and H_2O to give a total reaction volume of 25 ul.

Incubation followed, usually overnight, at room temperature and the reaction was terminated by the addition of 10 ul of Reaction Termination Buffer plus 90 ul of H_20 or Column buffer (section 2.7.1). Unincorporated radionucleotide was removed as in section 2.7.1 - i.e. Sephadex G-50 Column Chromatography.

2.7.3 Labelling Single Stranded M13 with α ^{32}PdCTP

The method used is essentially that of Hu and Messing (1982), enabling the preparation of probes from single stranded (s.s.) M13 subclones, subsequently for use in screening other M13 subclones on a hybridot filter (section 2.8.3)

Typically, in a 0.75 ml Eppendorf tube, 100 ng of S.S. M13 subclone DNA and 7 ng M13 labelling primer were annealed in buffer conditions of 10 mM Tris/HCl pH7.5, 6mM $MgCL_2$,60mM NaCl, 10 mM DTT, for 15 minutes at 65°C followed by gentle cooling to room temperature. Next dGTP, dATP, and dTTP were added to a final concentration of 50uM, along with 10uCi α ^{32}PdCTP (@ 10uCi/ul) and 4.5 units of Klenow fragment from DNA polymerase 1. The reaction was performed at 15°C for 90 minutes and terminated by adding EDTA (pH 8.3) to a final concentration of 25 mM.

Specific activity of the probe was determined by TCA precipitation (section 2.7.4) on a 1 ul aliquot of the sample, and was usually 1 x 10^6 c.p.m./ug or greater.

2.7.4 TCA Precipitation of α ^{32}PdCTP Labelled DNA

This was confined to use on s.s. M13 probes which had not been subjected to Sephadex G-50 column chromatography. Labelled DNA was precipitated on ice with 10% TCA and collected on cellulose acetate discs over a vacuum. After washing with 10% TCA and Ethanol, the discs were dried in an 80° vacuum oven prior to scintillation counting.

2.7.5 Scintillation Counting of Labelled DNA

This was performed on a Packard Prias Tricarb Scintillation Counter, model PL/PLD. Using white plastic scintillation vials with screw on caps, non-TCA precipitated liquid samples were counted in 5 mls of Liquiscint, while cellulose acetate discs containg TCA precipated samples were mixed with 5 mls POPOP (300 mg/l 1,4-Di-2(5-phenyloxazolyl)benzene in toluene) prior to counting. All sample vials were wiped with damp tissue to destroy interference from static electricity. Counting was usually over a 1 minute interval, with the machine giving a 90% reading efficiency for α ^{32}PdCTP.

2.8 Hybridisation of Probes to Nitrocellulose Filter

2.8.1 Southern Blotting

The method used was essentially that of Southern, 1975.

2.8.1.1 Transfer of DNA from Agarose Gel to Nitrocellulose Filter

The gel was placed in a glass dish and incubated at room temperature, with gentle shaking, for 1 hour in 2 changes of Denaturing solution (1.5 M NaCl, 0.5 M NaOH, 1mM EDTA) and then for 1½ - 2 hours in 3 changes of Neutralising solution (3 M NaCl, 0.5M Tris, 1mM EDTA, pH7.0 with HCl)

A deep sided plastic tray (approx 35 x 25 x 5 cms) was partially filled with 20 x SSC (3M NaCl, 0.3M Tri-Sodium Citrate, pH 7.0 with HCl) and a 'wick' of Whatman 3MM paper cut to size was suspended over the tank using a long glass plate. Once the wick was thoroughly soaked in 20 x SSC, the gel was gently slid into position on the wick (a small corner was removed from the gel to allow orientation). A nitrocellulose membrane filter, soaked for 2 minutes in distilled water, and then 20 minutes in 20 x SSC, was then carefully laid over the gel avoiding the appearance of bubbles beneath the filter. 6 filter size pieces of Whatman 3 mm paper and 2 nappyliners cut in half were then placed over the filter, and covered with a clean glass plate. Finally a 5 litre flask containing approx. 1 litre of water was positioned centrally on the glass plate, and the apparatus left overnight. Next day, after dismantling the apparatus, the wells and outline of the gel were traced on the filter, which was then baked for 2 hours at 80°C in a vacuum oven between two sheets of Whatman 3 MM paper.

2.8.1.2 Hybridisation of DNA Probe to the Nitrocellulose Filter

Firstly, the filter was prehybridised in 5 x SSC (section 2.8.1.1), 5 x Denhardt's solution (0.02% Ficoll, 0.02% Polyvinylpyrrolidone, 0.02% BSA) and 200 ug/ml herring sperm DNA, using 100 mls of this solution per whole filter, for 1 hour at 65°C in a sealed plastic bag placed in a flat based plastic container on a shaking water bath.

Hybridisation was carried out in a similar fashion, using 50 ml per whole filter of the following solution; 5 x SSC,2 x Denhardt's solution, 100 ug/ml herring sperm DNA plus the appropriate amount of probe solution (usually approx. 1 x 10^6 counts was sufficient for plasmid DNA blots, and 5 x 10^7 or more was used for genomic DNA blots). Before addition, double stranded DNA probes were denatured by boiling for 7 minutes followed by rapid cooling on ice. Hybridisation was performed overnight at 65°C.

Washes, each using 200 mls solution per whole filter were as follows; Once for 30 minutes in 2 x SSC, twice for 15 minutes in 1 x SSC and twice for 15 minutes in 0.1 x SSC. Finally the filter was air dried between 2 sheets of Whatman 3MM paper prior to autoradiography.

2.8.1.3 Autoradiography

The filter was mounted on white card backed by a glass plate. Small aliquots (1-2ul) of radiocative ink were applied assymetrically at either side of the origin and the filter assembly sealed in a polythene bag. Fuji X-ray film was then flash sensitised, placed over the filter and covered with a Dupont Cronex Lightning Plus Intensifying screen (Laskey and Mills, 1977). A second glass plate and 3 elastic bands were then used to hold everything in place, and exposure was then carried out inside 3 black plastic bags and a light proof box at -80°C. Films were developed in Ilford Phenisol developer for 5 minutes at room temperature, washed in cold water and then immersed for 2 minutes at room temperature in Kodak fixer. Finally, films were washed for 20 minutes in cold running water and then air dried.

2.8.2 Northern Blotting

2.8.2.1 Transfer of RNA from Gel to Nitrocellulose Filter

The method of Thomas, 1980, was used. After running the glyoxal gel (section 2.5.5) samples to be transferred were not stained with Acridine Orange (for staining purposes, duplicate samples were usually run on a separate section of the gel, which was then cut away). The blotting procedure was as for Southern Blots (section 2.8.1.1) but denaturing and neutralising was not necessary.

2.8.2.2 Hybridisation of Probe To The Filter

Using formamide in both prehybridisation and hybridisation solutions allowed incubation to be carried out at 42°C, which, although giving a slower annealing rate of probe to filter, meant less damage to the RNA bound to the filter.

Prehybridisation was for 4-6 hours, at 42°C, in 50% deionised formamide, 5 x Denhardt's solution, 5 x SSC (section 2.8.1.2) and 100 ug/ml herring sperm DNA at 100 ml per whole filter again in a sealed polythene bag in a box on a shaking water bath.

Hybridisation was performed overnight at 42°C, the denatured probe being added to 50% formamide, 2 x Denhardt's solution, 5 x SSC and 200 ug/ml herring sperm DNA in a volume of 50 ml per whole filter.

Washing was twice for 10 minutes in 2 x SSC, 0.1% SDS at 42°C, twice for 10 minutes in 0.1 x SSC, 0.1% SDS at 42°C, and finally a high stringency wash at 50°C in 0.1 x SSC, 0.1% SDS, using 100 mls of wash solution per whole filter.

The filter was then air dried and autoradiographed as in section 2.8.1.3.

2.8.3 Screening M13 Subclones by the Hybridot method

2.8.3.1 Filter Preparation

This technique facilitates the simultaneous screening of large numbers of M13 subclones (section 2.9.1).

The Nitrocellulose hybridot filter was prepared using Hybridot apparatus (Hybri-Dot^{TM},B.R.L.). The filter was cut to size, soaked for 2 minutes in water and then 20 minutes in 20 x SSC (section 2.8.1.1.), assembled in the apparatus and attached to a water powered vacuum pump according to accompanying instructions. A 1ul aliquot of each M13 subclone diluted in 39 ul of 20 x SSC was enough to produce duplicate filters. After loading (as detailed in instructions) each sample was washed twice with an equal volume of 20 x SSC. Finally the apparatus was dissembled and the filter dried under vacuum at 80°C.

2.8.3.2 Hybridisation of Single Stranded M13 Probe to the Filter

Single stranded M13 probe (section 2.7.3.) was hybridised to the filter using exactly the method detailed in section 2.8.2.2 and washing was performed to the highest stringency.

2.8.4 M13 Plaque Lifts Onto Nitrocellulose Filter

Using the method outlined in Maniatis et al., 1982, this was another way of screening large numbers of M13 subclones, with the potential advantage that only the required subclone plaques need be miniprepped (section 2.9.3). However this was not often a feasable method, if plaques were too densely spaced on the plate.

Gridded Nitrocellulose filters were numbered with biro, and then at room temperature, each laid onto a dish in direct contact with the plaques and with no air bubbles. The filter and underlying agar were pierced with a hole punch in an assymetric pattern for later realignment. After 60 seconds, the filter was removed with blunt ended forceps, and immersed, DNA side up, firstly in denaturing solution (section 2.8.1.1) for 60 seconds, and then in neutralising solution (section 2.8.1.1) for 5 minutes. Finally the filter was rinsed in 2 x

SSC (section 2.8.1.1), air dried between 2 sheets of Whatman 3MM paper, and baked in a vacuum oven for 2 hours at 80°C.

Hybridisation was then as detailed in section 2.8.1.2, using a denatured double stranded DNA probe.

Following autoradiography, a tracing of the filter hybridisation pattern was made onto clear polythene and by alignment with holes made on the original plate, desired plaques were identified and picked off for miniprepping (section 2.9.3).

2.9 DNA Sequencing

The Dideoxy chain termination method using single stranded 'phage M13 vectors, developed by Sanger et al., 1977 was used. The protocol employed was basically as outlined in the Amersham booklet 'M13 cloning and sequencing Handbook' (P1/129/84/10).

2.9.1 Construction of M13 Subclones

Two different strategies were employed in the subcloning step. In both cases a molecular ratio of 23:1, insert:vector, was used. M13 mp8, mp9, mp18, mp19 vectors (section 2.1. Table 2) were used exclusively.

2.9.1.1 Shotgun Cloning (Messing 1981)

The DNA was restricted, usually with an enzyme having a 4 bp recognition site, producing a range of small fragments, which were then ligated into suitably restricted M13 vector.

2.9.1.2 Direct (forced) Cloning

Used to subclone a specific DNA fragment, this could be achieved in two ways. Firstly by fragment isolation (section 2.6) followed by ligation into compatably cut M13, or secondly by restricting at unique sites encompassing the desired fragment ensuring that it be the only insert present in the resulting M13 subclones.

2.9.2 Transfection Into E.coli

2.9.2.1 Preparation of Competent Cells

An overnight culture of JM 101 (see Table 2) was prepared by inoculating 5 mls of 2 x YT broth (see Table 1) with JM101 grown on minimal agar. 50 ul of this overnight was then used to innoculate 50 mls of 1 x YT broth (Table 1) in a baffled flask. The cells were then grown at 37°C with constant shaking until the O.D.$_{660}$ reached 0.5 (approx. 4 hours). The cells were then spun at 4000g for 5 minutes (MSE Highspeed 18). The supernatant was discarded and the pellet resuspended in 40 mls of 0.1M Ca Cl_2 and then left on ice for 1 hour. This was followed by a second spin at 4000g for 5 mins, disposal of supernatant, and finally, resuspension of the pellet in 1 ml of 0.1M $CaCl_2$. The cells could be used immediately or stored for up to 24 hours on ice.

2.9.2.2 The Transfection Step

Firstly, 100 ul of competent cells were mixed with the M13: insert ligation and left on ice for 1 hour. The sample was then heat shocked by incubation at 42°C for 3 minutes, and 10ul of 0.1M IPTG, 50ul 2% Xgal(in Dimethyl Formamide) and 200ul of exponentially growing JM 101 cells were added and thoroughly mixed. This mixture was then combined with 3 mls of YT top layer agar (see Table 1) kept molten at 42°C, mixed, and poured evenly onto a plate of YT Agar (Table 1). Once set, the plate was incubated overnight at 37°C.

2.9.3 Miniprepping of Single Stranded M13 DNA

M13 recombinant plaques were selected by inactivation of the β -galactosidase marker and consequent lack of blue colouration (Messing 1983). Amplification of the recombinant DNA to levels sufficient to serve as a template for DNA sequencing was achieved by following the miniprep method outlined in the Amersham M13 cloning and sequencing handbook (P1/129/84/10):

2 mls of 2xYT broth was inoculated with 2 ul of a fresh JM101 overnight culture (section 2.9.2.1.) in a Universal bottle. A sterile cocktail stick, touched onto the surface of a selected plaque, was then introduced into the broth, followed by incubation overnight at 37°C.

1.5 mls of the culture was then transferred to a 1.5. ml Eppendorf tube and centrifuged at 12,000g for 4 minutes (MSE Microcentaur). Approx. 1.3mls of the resulting supernatant* was transferred to a fresh Eppendorf tube and the above spin repeated. 1 ml of this new supernatant was then combined with 200 ul of 20% Polyethylene Glycol (PEG), 2.5M NaCl in another fresh Eppendorf tube and vortex mixed. The sample was then left for at least 20 minutes at room temperature, followed by a 4 minute spin at 12,000g. All supernatant was then carefully removed and discarded, the pellet thoroughly resuspended in 100 ul TE buffer (section 2.3.1), Phenol extracted once with Phenol (section 2.3.1.) and Ethanol precipitated (section 2.3.2). Finally the sample was resuspended in 20 ul TE buffer and stored at -20°C.

* The pellet from this spin was kept for up to 2 months at 4°C in case the cells were needed for long term storage under glycerol (section 2.10.3.)

2.9.4 Screening of M13 Recombinants

2.9.4.1 Agarose Gel

1ul aliquots of each miniprep were run on a 0.7% Agarose gel against a standard sample of single stranded M13, allowing selection of recombinants by difference in size.

2.9.4.2 Hybridot screening

Up to 96 M13 minipreps could be screened simultaneously using the Hybridot method (section 2.8.3 details filter preparation, section 2.7.3 describes labelling of single stranded probes)

2.9.4.3 Plaque Lifts

As detailed in Section 2.8.4 this method avoids indiscriminate miniprepping of random M13 recombinant plaques, but could only be used when the density of plaques was fairly low.

2.9.4.4 Insert Orientation

A subclone containing an insert in one orientation could be used to screen other recombinants to show if any contained the same insert, but in the opposite orientation.

The method used was that outlined in Focus 1984. 1ul of the test subclone was reacted with 1ul of the known orientation subclone under the following conditions - 0.3M NaCl, 0.6% SDS, 0.02% Bromophenol blue, 5 mM EDTA pH8.0, and 12% Formamide, incubated at 65°C for 1 hour. This was then immediately loaded and run on a minigel, a positive result being indicated by a succession of bands, indicating annealing between the two subclones.

2.9.5 DNA Sequencing

The method was that described in the Amersham M13 cloning and sequencing handbook. The process consists of two stages - annealing of primer to the M13 template, followed by synthesis of radioactive fragments from the template.

2.9.5.1 The Annealing Reaction

5 ul of M13 template and 2 ul 17 mer primer were annealed in 10 mM Tris/HC1 pH 8.0, 5mM $MgCl_2$ for 5 minutes at 85°C followed by slow cooling to room temperature. The mixture was then spun briefly to the bottom of the tube.

2.9.5.2 The Sequencing Reaction

1 ul (8-10uCi) of α ^{35}S dATP and 1 unit of Klenow fragment were added to the annealed template and primer. Then, into 4 long (0.5ml) Eppendorfs, labelled A,C,G and T, was placed 2ul of the appropriate nucleotide mix (see Table 3). The tubes were then spun briefly, and 2.5 ul of template primer mix added to each and again spun briefly,starting the reaction. Incubation was for 20 minutes at 30°C, after which time 1 ul of dATP chase solution (table 3) was added to each tube, spun down, and incubation continued for a further 15 minutes at 30°C. Finally, 4 ul of formamide dye mix (96% deionized formamide, 0.1% Xylene Cyanol F.F., 0.1% Bromophenol blue, 20 mM EDTA pH 8.0) was added to each tube. Samples were then stored at -20°C prior to electrophoresis on polyacrylamide gel (section 2.9.6).

TABLE 3 Nucleotide Solutions for DNA Sequencing

Deoxy NTP mixes (A°, C°, G°, T°)

	A°	C°	G°	T°
0.5mM dCTP	20 ul	1 ul	20 ul	20 ul
0.5mM dGTP	20 ul	20 ul	1 ul	20 ul
0.5mM dTTP	20 ul	20 ul	20 ul	1 ul
TE buffer	20 ul	20 ul	20 ul	20 ul

Dideoxy NTPs-working solutions (concentrations occasionally modified)

dd ATP	0.125 mM
dd CTP	0.5 mM
dd TTP	1.0 mM
dd GTP	0.5 mM

Final Nucleotide mixes

Equal amounts of deoxy NTP and the corresponding dideoxy working solution were added to one another, mixed and stored at -80°C

dATP Chase Solution

dATP at a concentration of 0.5 mM

N.B. All dilutions made in double distilled deionized water.

2.9.6 Separation of Labelled Fragments Using Polyacrylamide Gel Electrophoresis

Two 6% Polyacrylamide gels were prepared. 2 pairs of glass plates (40 x 20 cm) were thoroughly cleaned with 1, 2-Dichloroethane, and one plate of each pair siliconised and polished with water and Ethanol on its inner side. 0.4 cm thick spacers were placed between the plates, and Scotch Electrical tape used to tape each pair together and give a watertight seal.

Next, 100 mls of gel solution (enough for two gels) was prepared: After deionizing a 38% Acrylamide, 2% Bis-acrylamide (N,N'-Methylenebisacrylamide) solution, 15 mls of this was mixed with 10 mls of 10 x strength TBE buffer (0.89 M Tris, 0.88 M Boric acid, 0.25 M EDTA). 50g Urea was added and dissolved (using gentle heat), water added to give a final volume of 100 mls, the solution filtered through a cellulose acetate filter, and degassed in a side arm flask. Gels were cast after adding 600 ul of freshly made 0.1M Ammonium Persulphate and 35 ul TEMED (N,N,N,N, Tetramethylene Diamine) and allowed to set for at least 1 hour. Electrophoresis was performed on vertical tanks in TBE buffer (0.089 M Tris, 0.088 M Boric Acid, 0.025 M EDTA) using Aluminium heat spreading plates. Prior to sample loading, gels were pre-electrophoresed for 1 hour at 1500 V, samples were then denatured for 5 minutes at 90°C, rapidly cooled on ice, and then duplicate 3ul aliquots loaded onto each gel using a fine glass capillary tube. The current was then turned on to approx. 50 mA (and 1500V), one gel run to the point where the Bromophenol blue dye just runs out of the bottom of the gel (usually approx. 2 hours) and the other for ½ hour after the Xylene Cyanol dye band had run out (approx. 5 hours).

2.9.7 Autoradiography

Firstly, the gels were dried onto Whatman 3 mm paper under vacuum overnight, and then exposed to X-ray film (unflashed) between glass plates inside 3 layers of black polythene for between 1 and 3 days. Films were developed as in section 2.8.1.3.

2.10 Transformation of Plasmid DNA into E.Coli

2.10.1 Transformation Step

Competents cells of JM 83 (see Table 2) were prepared using the method shown in section 2.9.2.1. 100 ul of the cells were mixed with approx 0.1 ug of plasmid DNA and left on ice for 30 minutes. The sample was then heat shocked for 5 minutes at 37°C, 1 ml of 1x YT broth (see Table 1) added, followed by a further 45-60 minutes incubation at 37°C.

100 ul aliquots of the transformation mixture were then spread on YT/Xgal/Amp plates (Table 1) and when dry, the plates incubated overnight at 37°C. Remaining transformation mixture was stored on ice for up to 48 hours. If required, this could be plated out by giving the solution a brief spin (MSE Microcentaur), discarding the supernatant, and spreading the pellet onto a fresh YT/Xgal/Amp plate.

2.10.2 Plasmid DNA Minipreps

The method used was that of Birnboim and Doly (1979).

A sterile toothpick was touched onto the colony on the plate, a streak made on a gridded reference plate (YT/Xgal/Amp), and the stick introduced into 10 mls of 1 x YT broth (Table 1) supplemented with 50 ul of 10 mg/ml Ampicillin in a McCartney bottle, and the cells grown up overnight at 37°C, on a rotary shaker. The reference plate was also incubated at 37°C for the same period, and then stored at 4°C for up to 2 months.

The culture broth was then spun for 10 minutes at 3000 rpm (MSE Centaur), the supernatent drained off and the pellet resuspended in 200 ul of solution 1 (freshly prepared; 2 mg/ml lysozyme, 50mM glucose, 10 mM EDTA, 25 mM Tris/HCl pH 8.0), transferred to a 1.5 ml Eppendorf, and left on ice for 30 minutes. 600 ul of solution 2 (freshly prepared; 0.2M NaOH, 1% SDS) were then mixed in, and the tube left a further 15 minutes on ice. Next, 450 ul of solution 3(3 M Sodium Acetate, pH 4.8 with glacial acetic acid) were added, mixed, and the sample put back on ice for 1 hour.

This was followed by a 5 minute spin at 12000g (MSE Microcentaur), 1100 ul of supernatant transferred to a fresh Eppendorf tube, 500 ul of Isopropanol added, and the DNA precipitated at -20°C for 30 minutes, and pelleted out by a 3 minute spin at 12,000g. The supernatant was discarded, and the pellet thoroughly resuspended in 200 ul of 0.1M Sodium Acetate /0.05 M Tris/HCl, pH 6.0. The DNA was then precipitated using 2 volumes of Ethanol at -20°C for 20 minutes, spun down and resuspended as above, and finally Ethanol precipitated again, to give a pellet which was dried down, dissolved in 100 ul of sterile water, its concentration determined (section 2.12) and stored at -80°C.

2.10.3 Longterm Storage of Bacteria Under Glycerol

A sterile loop was used to transfer cells from either plate or M13 pellet (section 2.9.3) to a small screwcap vial containing 1ml 80% glycerol and 1 ml 1 x YT broth (Table 1). Following thorough mixing the sample was stored at -80°C.

2.11 Extraction of Genomic DNA From Leaf Material

2.11.1 Genomic DNA Extraction

The method used was scaled down from that of Graham (1978).

4g of frozen leaf material (section 2.1.2) was weighed out, and ground to a very fine powder in a pestle and mortar using copious quantities of liquid Nitrogen. After transfer to a sterile, siliconised 100 ml beaker, the material was mixed with 10 mls of homogenizing buffer (0.1M NaCl, 0.025 M EDTA, pH8.0, 2% SDS, 0.1% Diethyl Pyrocarbonate), 5 mls of 5M Sodium Perchlorate, ½ volume of redistilled Phenol, ½ volume of Chloroform/Octanol (99:1), and shaken on ice for 20 minutes. The resulting suspension was centrifuged at 12,000g (Sorvall RC-5B) for 1 minute at 10°C in 30 ml Corex tubes. The supernatant was then removed and shaken with an equal volume of Chloroform/Octanol, and recentrifuged as above. The new supernatant was then combined with 2 volumes of -20°C Ethanol, the DNA spooled out, using a siliconised glass hook, and transferred to a preweighed Corex tube to be resuspended by shaking overnight on ice in 2 mls of resuspension buffer (50 mM Tris/HCl, 10mM EDTA, pH 8.0). Self-digested Pronase P (2 hours at 37°C) was then added to a final concentration of

500 ug/ml, and the sample incubated at 37°C for 3 hours.

The next stage was purification on a Caesium Chloride (CsCl) gradient. Firstly, the weight of the resuspended DNA solution was determined, and 30 ul of 10 mg/ml Et Br added per gram of solution. After reweighing, CsCl was added in the ratio of 0.94g per gram of DNA sample and dissolved. Centrifugation was then performed in quickseal tubes, at 50,000g for 36 hours using a Sorvall Ultra Centrifuge (Model OTD 65B). The resulting DNA band was visualised over a U.V. transilluminator, removed using a 19G2 gauge needle and a 10 ml syringe, and decolourised in a Corex tube by mixing with 7-8 changes of Isoamyl alcohol added in equal volume.

2.11.2 Dialysis of the DNA Solution

An appropriate length of dialysis tubing was boiled in distilled water with a couple of drops of 0.2M EDTA pH8.0 for 15 minutes, then in pure distilled water, again for 15 minutes. One end of the tubing was sealed with a Medi-clip. Dialysis was then performed in 1 litre of resuspension buffer (section 2.11.1) changed 3 times, over 48 hours, with constant magnetic stirring. Finally, the DNA was Ethanol precipitated (section 2.3.2) resuspended in resuspension buffer, and the concentration of DNA determined using Spectrophotometry (Section 2.12).

2.12 Determination of DNA Concentration Using Spectrophotometry

Estimation of the DNA Concentration of a solution was determined using a Pye-Unicam SP8-150 u.v.-vis scanning spectrophotometer. At

$O.D._{256}$ a reading of 0.02 corresponds to a DNA concentration of 1 ug/ul (using 1 ml quartz cuvettes in a 1 cm light path). Contamination of samples with phenol or protein was assessed by scanning the absorbance of the solution between $\lambda_{300n.m}$ to $\lambda_{200n.m.}$

2.13 Statistical Methods and Tests

2.13.1 Construction of Nucleic Acid sequence homology profiles and determination of the level at which homology becomes significant

The nucleotide sequences to be compared were aligned for maximum homology using the programme 'NUCALN' on an IBM Personal Computer. Percentage homology was then determined in 20 base segments, overlapping each time by 10 bases and the resulting data plotted against the length of the sequences on a graph, giving an overall profile of areas of strongest sequence conservation.

A threshold level at which homology becomes significant was determined as follows: Over the 20 base stretch compared, each base has a 0.25(¼) chance of matching and a 0.75(¾) chance of not matching. Consequently the probability of all 20 bases matching = $(0.25)^{20}$ = 9.536×10^{-7}, or 1 chance in approx 1,050,000, which, assuming a significance value of $p < 0.05$ (p = probability) causes a rejection of the Null Hypothesis.

In fact, the level of homology which gives a p value closest to 0.05 was found to be 40%. At this level there are 8 matches and 12 non-matches. Hence the equation $(p + q)^{20}$ was expanded, the term relating to p^8,q^{12} was found and multiplied out, giving a value of

0.06. Hence a line was drawn on the homology profile at 40%, indicating that any part of the plot occurring above this line shows significant sequence homology.

2.13.2 Calculation of Codon Adaption Index (CA1) Values for Storage Protein Genes.

The method used was taken from that of Sharp and Li, 1987. A table of composite codon usage amongst all legume storage proteins thus far sequenced was compiled (data not presented).

Next, a reference table of relative synonymous codon usage (RSCU) values was constructed from these genes (Table 4), where

$$\mathrm{RSCU} = \frac{Xij}{\frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij}}$$

X_{ij} = number of occurrences of the jth codon for the ith amino acid, and n_i is the number (from 1 to 6) of alternative codons for the ith amino acid.

The relative adaptiveness of a codon W_{ij} is then the frequency of use of that codon compared to the frequency of the optimal for that amino acid.

$$Wij = \mathrm{RSCU}_{ij}/\mathrm{RSCU}_{i\ max} = X_{ij}/X_{i\ max}$$

Where $\mathrm{RSCU}_{i\ max}$ and $X_{i\ max}$ are the RSCU and X values for the most frequently used codon for the ith amino acid. W_{ij} values for the genes are also given in Table 4.

The Codon Adaption Index (CAI) for a given gene is then calculated as the geometric mean of the RSCU values corresponding to each of the codons used in that gene, divided by the maximum possible CAI for a gene of the same amino acid composition, i.e.

$$\mathrm{CAI} = \mathrm{CAI}_{obs}/\mathrm{CAI}_{max}$$

where

$$\mathrm{CAI}_{obs} = \left(\prod_{K=1}^{L} \mathrm{RSCU}_{K}\right)^{1/L}$$

$$\mathrm{CAI}_{max} = \left(\prod_{K=1}^{L} \mathrm{RSCU}_{Kmax}\right)^{1/L}$$

and where RSCU_{K} is the RSCU value for the Kth codon in the gene, RSCU_{kmax} is the maximum RSCU value for the amino acid encoded by the Kth codon in the gene, and L is the number of codons in the gene.

Additionally any X_{ij} of value zero (ie the codon is never used in the reference set) is assigned a value of 0.5, to avoid the CAI value for a gene in which this codon appears becoming zero also. Secondly the number of ATG and TGG codons are not included in the final L value, since their RSCU values are fixed at 1.0 and so don't contribute to the CAI.

There is no intrinsic effect of gene length (L codons) on CAI, however, CAI values for short genes may be more variable due to sampling effects.

TABLE 4 R.S.C.U.ij and Wij Values for Legume Seed Storage Protein Genes

		RSCU ij	Wij			RSCU ij	Wij			RSCU ij	Wij
F-Phe	TTT	0.852	0.742	A-ALa	GCA	1.354	0.937	R-Arg	CGA	0.545	0.262
	TTC	1.148	1.000		GCG	0.217	0.150		CGG	0.185	0.089
					GCC	0.984	0.681		CGC	0.906	0.673
					GCT	1.445	1.000		CGT	0.885	0.426
L-Leu	CTA	0.672	0.371						AGA	2.079	1.000
	CTG	0.596	0.329						AGG	1.400	0.673
	CTC	1.285	0.709	T-Tyr	TAC	1.125	1.000				
	CTT	1.813	1.000		TAT	0.875	0.777				
	TTA	0.476	0.263					S-Ser	TCA	0.987	0.634
	TTG	1.157	0.638						TCG	0.237	0.152
				H-His	CAC	1.204	1.000		TCC	0.797	0.512
					CAT	0.796	0.661		TCT	1.557	1.000
I-Ile	ATA	0.761	0.572						AGC	1.376	0.883
	ATT	1.331	1.000						AGT	1.044	0.670
	ATC	0.908	0.682	Q-Gln	CAA	1.277	1.000				
					CAG	0.723	0.566				
								G-Gly	GGA	1.639	1.000
M-Met	ATG	1.000	1.000						GGG	0.489	0.298
				N-Asn	AAC	1.259	1.000		GGC	0.722	0.440
					AAT	0.741	0.589		GGT	1.149	0.701
V-VAL	GTA	0.619	0.390								
	GTG	1.586	1.000								
	GTC	0.418	0.264	K-Lys	AAA	0.993	0.986	W-Trp	TGG	1.000	1.000
	GTT	1.377	0.868		AAG	1.007	1.000				
P-Pro	CCA	1.521	1.000	D-Asp	GAC	0.957	0.917				
	CCG	0.281	0.185		GAT	1.043	1.000				
	CCC	0.688	0.452								
	CCT	1.511	0.993								
				E-Glu	GAA	1.053	1.000				
					GAG	0.947	0.899				
T-Thr	ACA	1.157	0.853								
	ACG	0.314	0.231								
	ACC	1.357	1.000	C-Cys	TGC	1.100	1.000				
	ACT	1.171	0.863		TGT	0.900	0.818				

2.13.3 Compilation of Dendrograms for Multi-Gene Families

All amino acid sequences available for members of a given multigene family were aligned, by hand (Computer Algorithms proved unreliable for this task), to give maximum homology. Sequences were then compared pairwise, the number of residues not matching counted up, and this figure divided by the number of residues compared (method from Miyata et al., 1980). Once completed this procedure produced a matrix of values of dissimilarity encompassing all members of the gene family.

Next, Single-Link Classification was carried out. The method was developed by Sneath, 1957, and essentially it seeks to fuse into clusters the least dissimilar individuals, or groups of individuals, at each stage, until ultimately all individuals are contained within one group. The vertical height of the nodes in the dendogram is proportional to the dissimilarity coefficient (DC) value between the individuals, or groups of individuals, fusing at each node, and the method uniquely produces a classification for which the sum of the DC-values is minimised.

The procedure:-

N.B. Ignore trivial 0.0. values for D_{ij} in all cases (where D_{ij} = DC value for genes i and j.

i) Set cycle number (C) to 0.

ii) Find the smallest D_{ij} in the matrix not yet crossed out or boxed. Increment C by 1. Box the D_{ij} value and label it with the value of C.

iii) For all other individuals k ($k \neq i$ or j, and k not clustered with i or j), k = 1, n;

If both D_{ik} and D_{jk} are not yet crossed out, cross out the larger of the two, and proceed to next k.

Otherwise, when either or both of D_{ik} and D_{jk} are crossed out; if k belongs to a cluster, and any m k (where m are the other individuals with which it is clustered), then proceed to the next k. Else, for each crossed out value D_{ak} (where a is i or j), seek an alternative DC-value by examining in turn all D_{am}, D_{lk} and D_{lm} (where m are the other individuals with which k is clustered, and l are other individuals with which a is clustered). When two values are located, cross out the larger of the two and proceed to the next k.

iv) Record the link between i and j (conveniently be sketching a dendogram as the analysis proceeds). Unless all individuals are now linked return to step (ii) until linkage is complete.

CHAPTER 3 - RESULTS

3.1 Nucleotide Sequencing of Legumin Gene D (*Leg* D), and its flanking regions

3.1.1. Restriction mapping of *Leg* D

λ Leg 1 is a genomic clone that has been isolated by hybridising a legumin cDNA probe, pDUB3, (Croy et al., 1982) to a bacteriophage λ gene library containing fragments produced by partial digestion of pea leaf DNA with Eco R1.

Digestion of the single, 13.5 Kb, Eco R1 fragment contained in λ Leg 1, with the restriction enzymes Hind III and Xbal, followed by hybridisation of pDUB3, identified the presence of two legumin genes; *Leg* A (whose complete coding, 5' and 3' flanking sequences have previously been reported (Lycett et al., 1984) and *Leg* D 1.3 Kb away from the 3' end of *Leg* A (fig.1).

The 1.5 Kb Hind III fragment corresponding to *Leg* D, and the 0.8 Kb Hind III fragment immediately 5' to it, were then subcloned into the plasmid vector pUC8, to give the subclones 6.01 and 6.03 respectively (see fig.1). These were then restriction mapped by digestion with a variety of 6 and 4 bp restriction enzymes to give the completed map shown in fig. 2.

Fig.1 Partial restriction map of the pea genomic clone λ Leg 1. Bars represent areas corresponding to the coding sequences plus introns of *Leg* A and *Leg* D. Arrows a and b indicate fragments used as probes on pea genomic DNA.6.01. 6.03 and 6.06 indicate subclones used for sequencing. Restriction sites : B = Bam H1, E = Eco R1, H = Hind III, S = Sst I and X = Xho I.

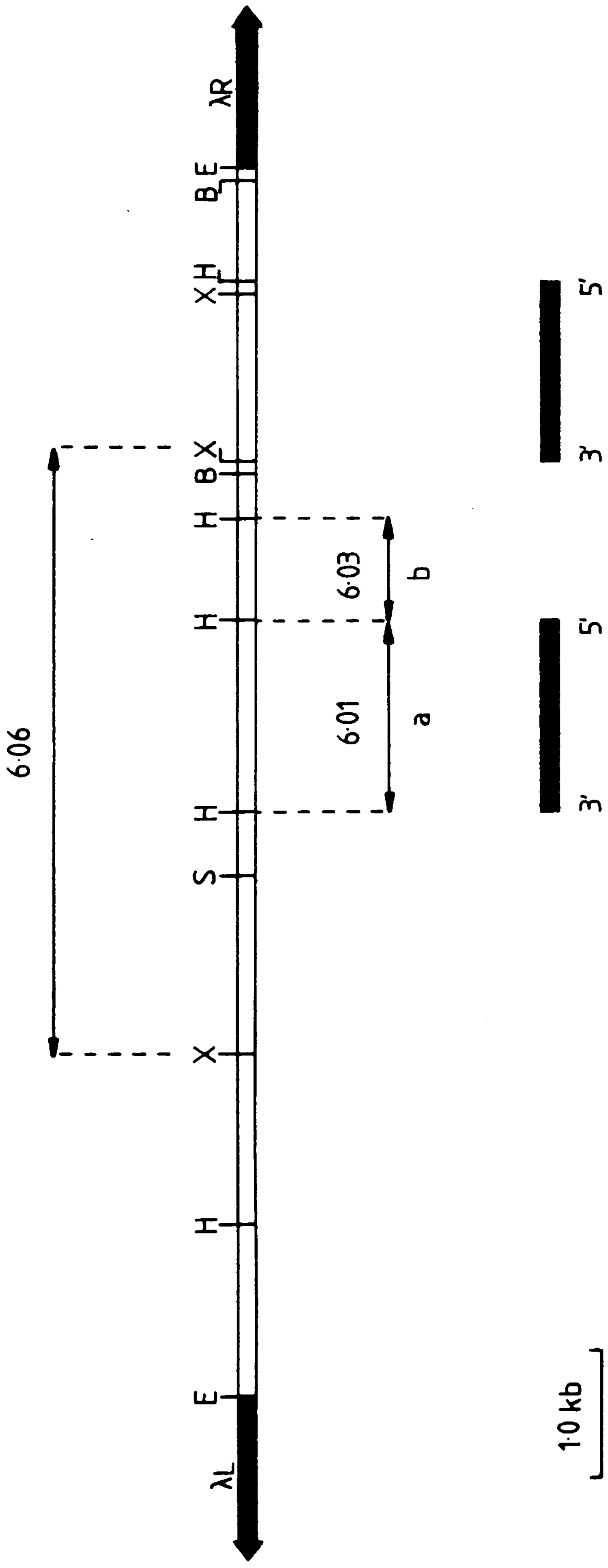
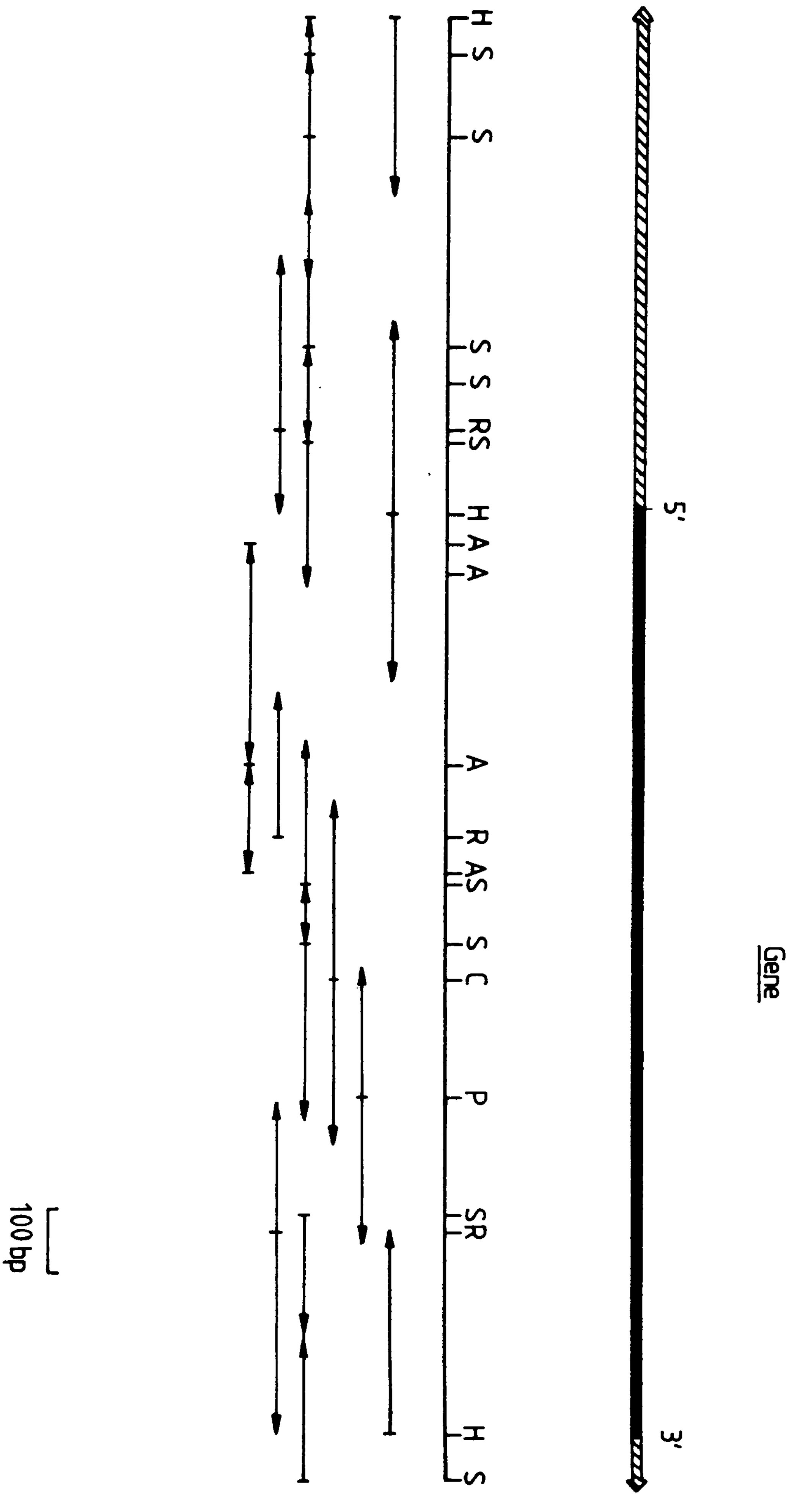

λL
E
H
X
S
H
H
H
B
X
X
H
B
E
λR
6·06
6·01
6·03
a
b
3'
LegD
5'
3'
LegA
5'
1·0 kb

Fig.2 Restriction and sequencing maps of *Leg* D and flanking sequences. Restriction sites: A = Alu I, C = Acc I, H = Hind III, P = Pst I, R = Rsa I, S = Sau 3A. Arrows indicate direction and length of sequence obtained from each M13 subclone. Duplicated sequencing runs carried out to check sequences are not shown. ■ = *Leg* D coding and intron sequence, ▨ = *Leg* D flanking sequence.


Gene
5'
3'
H
S
S
S
S
RS
H
A
A
A
R
AS
S
C
P
SR
H
S
100bp

3.1.2. The Complete Nucleotide Sequence of *Leg D* and its flanking regions

Nucleotide sequence corresponding to all the coding region and 800 bp of 5' flanking sequence was obtained from M13 subclones containing fragments of either 6.01 or 6.03 digested with Alu I, Acc I, Hind III, Pst I, Rsa I or Sau 3AI. The 0.5 Kb Hind III/Sst I fragment from pUC8 subclone 6.06 (see fig. 1) was then subcloned into M13 enabling the determination of 204 bp of 3' flanking sequence. The M13 dideoxy chain termination method of Sanger was used throughout to determine nucleotide sequence (see section 2.9.5). The exact sequencing strategy employed is included in fig. 2, and is indicated by arrows showing the direction of nucleotide sequence obtained from each M13 subclone.

After alignment for maximum homology the coding sequences of both *Leg* D and *Leg* A, along with their flanking regions are shown in fig.3.

Fig.3 Complete nucleotide sequence of *Leg* D and flanking regions. The sequence of *Leg* A(Lycett et al., 1984) is also given over the regions where the genes are homologous. The complete amino acid sequence of *Leg* A is given; residues differing in *Leg* D are given above the sequence. Stop codons are indicated. Δ = base deleted in *Leg* D; + = extra base in *Leg* D.

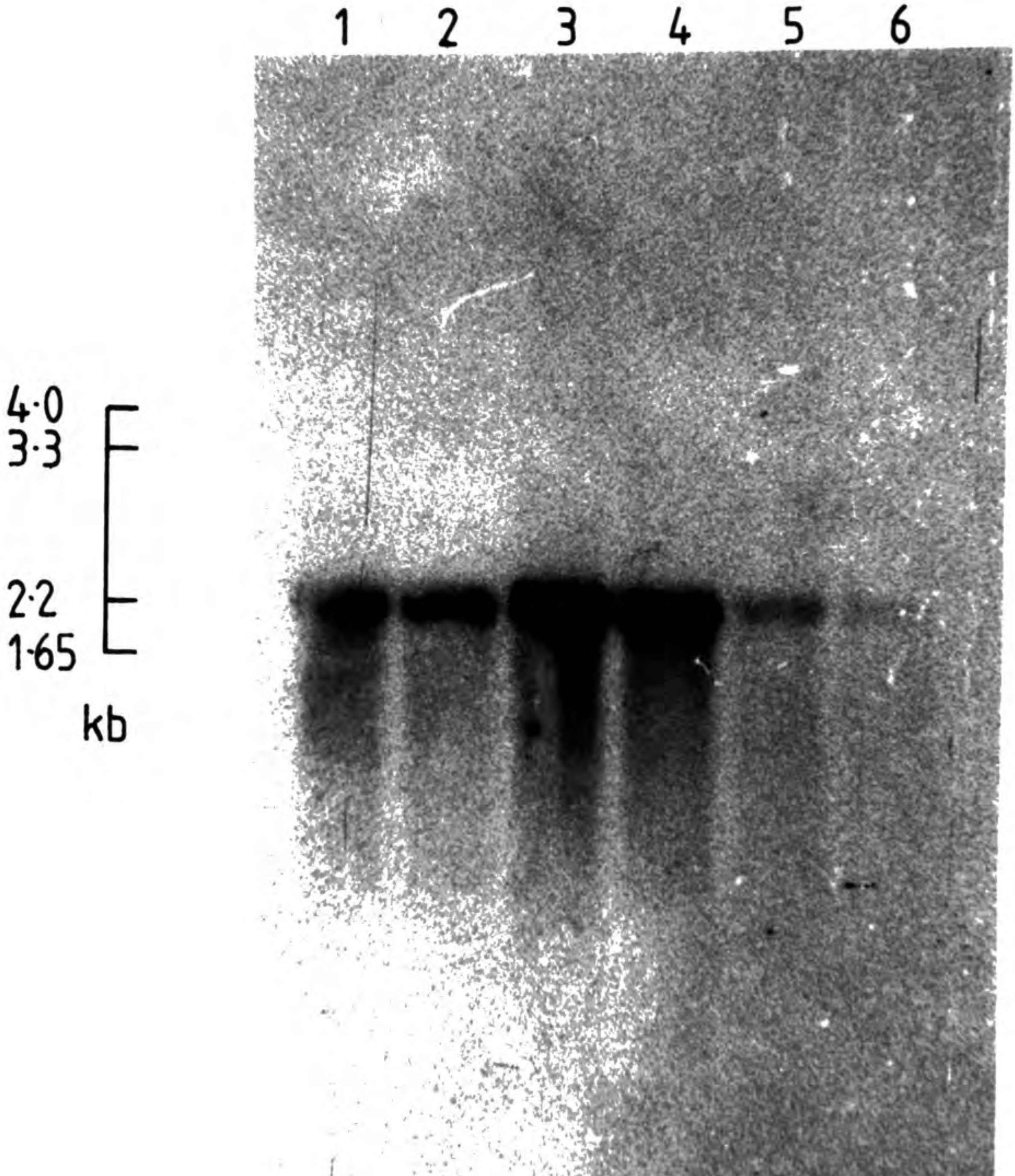
```
LegA 3′ flanking sequence / LegD .........(2488) AAGCTTATCCCTTCTTAGCATCAAGGTGTTGGCAAAGGATAAGCAACAAACACTTTTAGATCAATAAAAGTAGGTG
 -672


LegD AAGATTGAAACAAAATCTCATAGGTAGTTATGTTGTTTATCACATGAGAAAAGGCAATTTATTAATGCCGTGAGTGACATGGCATAAAGCAAAGGAACCAAACCGTTTTGAAAGTCTGAG -552


LegD GGAAAGCAGGATCTAATAACAATCAATAGGTGTCTATGAAAACTCTCAACAACCAGAATTTATTGTGAAGTTTCTACACAATTCTGGTGTAAAATGGTGAAGGAGATATAAAAGGAAGAC -432


LegD ATATTAAGTTACACTCCATTATCACTTTGTAGAAACAAATTTAACAAAAACTAGAAATGCAAGATTTAATGCCAAACTTATTTTTCATAAAGATTGCTCAAGTAAATTTAGAAAAGTTGT -312


LegD CAACAATAGTAAGGAAAATTGATAACTTGTAAGGAAATGTTGCTAATAGACCCCAAATGTCTATGCTATGAGAGAGATCAACACATTTAAGTGTCAGTAGGAGAAGATAATTTTGTTGGT -192
     .............................................................................(       AGGA /  CAAT BOX)..................
LegD TAGAAAATGACAATGATCACGAATGATGCAGTAATGAGACATTGTGAGGTGTAATGCAGAATACTTACATAGCCATGCAAGATGAAGAATGTCTAATGTACAGGAACCCATGCATACTCT -72
LegA GGTAATGGAGATGATGAAGCCATTAGCCACCTCCTCTATCAGACATAGGTGTAAAGCATTATGCTTCCATAGCCATGCAAGCTGCAGAATGTCCAATTCTCAACATCCCA--------C -73
     .............................................................................(       AGGA /  CAAT BOX)..................

     .......................................(TATA BOX).........................1..........................Start(     T
LegD CTGATCTGACGCGTCCCTCCTACACTCTACTCTCCTCCCCTATAAATATCAATGCCAAATTAAGGTTCTCCGCATCCCAAACA-------TATATTCTATCCAACTATGGCTACTAAGCT 40
LegA TTTCAATGACGTGTCC-ACCTTCACCACCCTCTCTTCTCATATAAATTACCACTTCTCATTAAGGTTCTCCGCATCACAACCAACATTCTCTTAGTATCTCTCTTCATGGCT---AAGCT 44
     .......................................(TATA BOX)..................(Start^mRNA)......................Start(M  A  -  K  L



                      L              F  S  S                       S Och R                          N        G           C
LegD TGCACTCTCTCTTTCCCTTTGTTTTCTACTTTTTAGTAGCTGCTTTGCTCTCAGAGAGCAGTCTTAACGAAATGAGTGTCAGCTGGAACGCCTCAATGCCCTCGGACCTGATAATTGTAT 160
LegA TGCACTTTCTCTTTCATTCTGTTTTCTACTTTTGGGCGGCTGTTTTGCTTTGAGAGAACAGCCACAGCAAAATGAGTGCCAGCTAGAACGCCTCGATGCCCTCGAGCCTGATAACCGTAT 164
A.A.   A  L  S  L  S  F  C  F  L  L  L  G  G  C  F  A  L  R  E  Q  P  Q  Q  N  E  C  Q  L  E  R  L  D  A  L  E  P  D  N  R  I

                  Umb    V                       R              V        T  L        T           P        F
LegD AGAGTCGGAAGGCTGACTGGTTGAGACTTGGAATCCAAACAACAGGCAATTCCGATGTGTGGGTGTGACCCTCTCTCGTACTACGCTTCAACCCAATGCCTTTCGCAGACCTTACTATTC 280
LegA AGAATCGGAAGGTGGGCTCATTGAGACTTGGAATCCCAACAACAAGCAATTCCGATGTGCTGGTGTGGCAATCTCTCGTGCTACCCTTCAACGCAACGCCCTTCGCAGACCTTACTACTC 284
A.A.   E  S  E  G  G  L  I  E  T  W  N  P  N  N  K  Q  F  R  C  A  G  V  A  I  S  R  A  T  L  Q  R  N  A  L  R  R  P  Y  Y  S

                   K     Y           (--------------------------------------------IVS-1-------------------------------------)
LegD CAATGCTCCCCAAAAAAATTTACATCCAACAAGGTTCGTTACTTTGATATTCTTTCAATGTCTTTACTTACATTACAGACCATACATATTTACTATTTT----TACTACATCAATTACTAG 396
LegA CAATGCTCCCCAAGAAATTTTCATCCAACAAGGTTACTTATTTTGATCTTATACCAACTTCTTTACGTACATTACATGCATATTAGCATACTATTAGTGTTCTACTATACCAATTACAAG 404
A.A.   N  A  P  Q  E  I  F  I  Q  Q  (--------------------------------------------IVS-1-------------------------------------)

                       I                                   L                 R                       S
LegD GTAATGGATATTTTGGCATTGTATTCCCTGGTTGTCCTGAGACCTTTGAGGAGCTACAAGAATCTGAACAAAGAGAGGGACGCAGGTATAGAGACAGCCACCAAAAAGTTAACCGATTCA 516
LegA GTAATGGATATTTTGGCATGGTATTCCCCGGTTGTCCTGAGACCTTTGAAGAGCCACAAGAATCTGAACAAGGAGAGGGACGCAGGTACAGAGACAGACATCAAAAGGTTAACCGATTCA 524
A.A. G  N  G  Y  F  G  M  V  F  P  G  C  P  E  T  F  E  E  P  Q  E  S  E  Q  G  E  G  R  R  Y  R  D  R  H  Q  K  V  N  R  F

                                S     V                                            I           E  T  G      Δ
LegD GAGAGGGTGATATCATTGCAGTTCCTAGTGGTGTTGTATTTTGGATGTACAATGACCAAGACACTCCAGTTATTGCCATTTCTCTTACTGAAACGGGTAGCTCC-ATAACCAGCTTGATC 635
LegA GAGAGGGTGATATCATTGCAGTTCCTACTGGTATTGTATTTTGGATGTACAACGACCAAGACACTCCAGTTATTGCCGTCTCTCTTACTGACATTAGAAGCTCCAATAACCAGCTTGATC 644
A.A. R  E  G  D  I  I  A  V  P  T  G  I  V  F  W  M  Y  N  D  Q  D  T  P  V  I  A  V  S  L  T  D  I  R  S  S  N  N  Q  L  D

                 (--------------------------------------------IVS-2----------------------------------------------------------
LegD AGATGCCTAGGGTGAGAACTAAGCATAATTAAACTTCCTGTATAATATATTAGATAATAATAAGCTAAATAATGTCCAAATCAGTAACATAGATCGGTAATTGTTTGATTTTGACTGAAT 755
LegA AGATGCCTAGGGTGAGCACTGAGCATAATTAAACTTCCCATATAAGATAATATGTTGTCCAAAACAGTAACATAGATTCTATCTATCTATGTTTCACAGAGATTCTATCTTGCTGGGAAC 764
A.A. Q  M  P  R (----------------------------------------------IVS-2------------------------------------)R  F  Y  L  A  G  N
```

```
     ----
LegD TATT----------------------------------------DELETION------------------------------------------------------------------TG 761
LegA CACGAGCAAGAGTTTCTACAATACCAGCATCAACAAGGAGGAAAGCAAGAACAAGAAAATGAAGGCAACAACATTTTCAGTGGCTTCAAGAGGGATTACTTGGAAGATGCTTTCAACGTG 884
A.A. H  E  Q  E  F  L  Q  Y  Q  H  Q  Q  G  G  K  Q  E  Q  E  N  E  G  N  N  I  F  S  G  F  K  R  D  F  L  E  D  A  F  N  V


     K   R              K                                                                                    R      P  H
LegD AAGAAGCATATAGTAGACAAACTCCAAGGCAGGAACAAAGATGAGGAGAAGGGAGCCATTGTCAAAGTTAAAGGTGGTCTCAGCATCATAAGCCCACCTGAGAGACAACCACATCACCAG 881
LegA AACAGGCATATAGTAGACAGACTTCAAGGCAGGAATGAAGACGAAGAGAAGGGAGCCATTGTCAAAGTGAAAGGTGGACTCAGCATCATAAGCCCACCCGAGAAGCAAGCGCGCCACCAG 1004

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


     K                                                                                                          C
LegD AAAGGCAGCAGACA------------------------------------DELETION----------------------AGATGAAGATGAAGAGAGGCAGCCGTGTCATCAAAGGAGA 935
LegA AGAGGCAGCAGACAAGAGGAAGATGAAGATGAAGAGAAGCAGCCGCGCCACCAGAGAGGCAGCAGACAAGAGGAAGAGGAAGATGAAGATGAAGAGAGGCAGCCGCGCCATCAAAGGAGA 1124
A.A. R  G  S  R  Q  E  E  D  E  D  E  E  K  Q  P  R  H  Q  R  G  S  R  Q  E  E  E  E  D  E  D  E  E  R  Q  P  R  H  Q  R  R


                              E            H  C                         W  H                          I                      Q
LegD AGAGGAGAGGAGGAAGAAGAAGACGAGAAAGAGCGCCACTGCAGCCAAAAAGGCAAAAGCAGATGGCATGGAGACAATGGGCTTGAGGAAACCATTTGCACGGCCAAACTTCGGCAGAAC 1055
LegA AGAGGAGAGGAGGAAGAAGAAGACAAGAAAGAGCGCGGCGGCAGCCAAAAAGGCAAAAGCAGAAGGCAAGGAGACAATGGGCTTGAGGAAACAGTTTGCACTGCTAAACTTCGATTGAAC 1244
A.A. R  G  E  E  E  E  E  D  K  K  E  R  G  G  S  Q  K  G  K  S  R  R  Q  G  D  N  G  L  E  E  T  V  C  T  A  K  L  R  L  N


           S               N              V                                           L  F                 I
LegD ATTGGTTCATCTTCATCACCTAACATCTACAACCCTGTTGCTGGTAGAATCAAAACTGTTACCAGCCTTGACCTCCCACTTTTCAGATGGCTCAAACTAATTGCTGAGCATGGATCTCTC 1175
LegA ATTGGCCCGTCTTCATCACCAGACATCTACAACCCTGAAGCTGGTAGAATCAAAACTGTTACCAGCCTGGACCTCCCAGTTCTCAGGTGGCTCAAACTAAGTGCTGAGCATGGATCTCTC 1364
A.A. I  G  P  S  S  S  P  D  I  Y  N  P  E  A  G  R  I  K  T  V  T  S  L  D  L  P  V  L  R  W  L  K  L  S  A  E  H  G  S  L


          <-------------------------------------------------IVS-3-----------------------------------------------)
LegD CACAAAGTACCATCATTTCTCTTTCTTTCTTTTTTCTCCTTTATTACTTTT-------AATTTATTTTCCATGACTTAATTCTATGTCAAACAATTTTCATACAGAATGCTATGTTCGTG 1288
LegA CACAAAGTATGTTTTTTCATCATTTAATTTGTTTTTCCATGAATCAATTTCATGTCGAACTATGTGTTGGAGAATAATAGCTAACTCATTACAATCTTCATACAGAATGCTATGTTTGTG 1484
A.A. H  K <-------------------------------------------------IVS-3-----------------------------------------------) N  A  M  F  V


                              C  V        T                           I                           S
LegD CCTCACTACAACCTCAACGCAAACTGCGTAATATACACACTGAAAGGACGTGCAAGGCTACAAATTGTGAACTGCAATGGCAACACTGTGTCTGATGGA--------------------- 1387
LegA CCTCACTACAACCTGAATGCAAACAGTATAATATACGCATTGAAGGGACGTGCAAGGCTACAAGTAGTGAACTGCAATGGCAACACCGTGTTTGATGGAAAGCTAGAAGCCGGACGTGCA 1604
A.A. P  H  Y  N  L  N  A  N  S  I  I  Y  A  L  K  G  R  A  R  L  Q  V  V  N  C  N  G  N  T  V  F  A  G  K  L  E  A  G  R  A


                                                                                      + K     A                             T
LegD -----------------------------------DELETION-------------------------------------GAAAGCTGCAATCGCCAGGCTTGCAGGGACATCCTCCACG 1427
LegA TTGACAGTGCCACAAAACTATGCTGTGGCTGCAAAGTCACTAAGCGACAGGTTCTCATATGTAGCATTCAAGACCAATGATAGAGCTGGTATTGCAAGACTTGCAGGGACATCATCAGTT 1724
A.A. L  T  V  P  Q  N  Y  A  V  A  A  K  S  L  S  D  R  F  S  Y  V  A  F  K  T  N  D  R  A  G  I  A  R  L  A  G  T  S  S  V


     L     A  M      V        I                           S              V                                   P     E
LegD CTAAATGCTATGCCAGTGGATGTAATAGCCGCTACATTCAACTTGCAGAGGAGTGAGGCTAGGCAAGTGAAGTCCAACAATCCTTTCAAATTTCTAGTTCCCCCTCGTGAGTCAGAAAAC 1547
LegA ATAAATAATCTGCCGTTGGATGTGGTTGCAGCTACATTCAACCTGCAGAGGAATGAGGCAAGGCAGCTCAAGTCCAACAATCCCTTCAAATTTCTAGTTCCAGCTCGTCAGTCTGAGAAC 1844
A.A. I  N  N  L  P  L  D  V  V  A  A  T  F  N  L  Q  R  N  E  A  R  Q  L  K  S  N  N  P  F  K  F  L  V  P  A  R  Q  S  E  N


     K  A  S  A Amb)Stop...................................................vPoly A Site1......vPoly A Site2..............
LegD AAAGCTTCTGCTTAGAAACAAACACTGCTTCGAAGCCTTTTTGTTTGAGAGACATGTATCCCACCCCAACTGGTAATAATAAAGATACTTATGAATAAAAAAAGGTTTGGTTTCCTTTTG 1667
LegA AGAGCTTCGGCTTAGATTTCGCACCAAATCAATGAAAGTAATGAATAAGAAAACTAAGGCTTAGATGCCTTTGTTACTTGTGTAAAATAACTCGAGTCATGTACCTTTTTGCGGAAACAG 1964
     R  A  S  A Amb)Stop......................^Poly A Site1.....................................................................


     .............................................vPoly A Site3...........................................................
LegD AAACTATGTATAAGTACTATCTGAGTCACGAACCTTTTCGAAAGTGAATAAAAGTAAAAGTGAAATTGAGCATACAAGTTTATTTTGATC (1757).......................
LegA AATAAATAAAAGGTAAAATTTCAGTGCTCTATGCTTTTCTACTCCAAGTTATAACCAGATGATATATATAACAATCACAATAAATAAATGTGAGTAAAAAAATATTGAAGAAAAATGATG 2084
     ^Poly A Site2...........<End^mRNA>........................................^Poly A Site3..........................
```

3.1.3. Hybridisation of *Leg* D to mRNA

A Northern blot (see section 2.8.2.1.) was prepared using pea total RNA isolated from cotyledons 12, 14 and 18 d.a.f. (days after flowering). The 1.5 Kb Hind III fragment corresponding to *Leg* D Coding Sequence (see fig.2) was then hybridised to the blot to give the autoradiograph shown in fig.4, where hybridisation to an mRNA species of approx 2000 bases appears to have occurred, corresponding to a predicted size of 1800 bases for legumin mRNA. No additional bands which might correspond to *Leg* D message (predicted size approx. 1400-1600 bases) were obtained. Intensities of the bands increased relative to the increased number of d.a.f. at which the samples were harvested, consistent with previous estimations (Gatehouse et al., 1982), also indicating that the only mRNA species detected was that of *Leg* A. In order to confirm this theory, the 0.5 Kb Hind III/Sst I fragment from 6.06 (see fig.1) was hybridised to a fresh Northern blot of pea cotyledon total RNA. This fragment corresponds to 3' flanking sequence specific only to *Leg* D , and its failure to hybridise to the RNA confirmed the absence of *Leg* D transcripts. (result not presented).

Fig.4. Hybridisation of *Leg* D (Hind III fragment a, fig.1) to total RNA prepared from pea cotyledons. Tracks 1, 2 : 10, 5 ug (respectively) RNA from cotyledons 14 d.a.f. (days after flowering); tracks 3, 4 : 10, 5 ug RNA from cotyledons 18 d.a.f.; tracks 5, 6 : 10, 5 ug RNA from cotyledons 12 d.a.f.


1
2
3
4
5
6
4·0
3·3
2·2
1·65
kb

3.1.4. Hybridisation of Leg D to Genomic DNA

The Hind III fragments from the pUC8 subclones 6.01 and 6.03 (see fig. 1) corresponding to *Leg* D coding sequence and 5' flanking sequence respectively were hybridised individually to two identical Southern blots (section 2.8.1) each containing pea leaf genomic DNA digested with the same 4 restriction enzymes (see section 2.4.1.2.), namely Bam H1, Taq 1, Hind III and Eco RI - see fig. 5 for the gel photograph. The coding sequence probe produced the result shown in fig.6, with intense bands at 12.5 Kb in the Eco RI digest, 2.4 Kb in Hind III, 1.1 Kb in Taq 1 and 3.6 Kb in Bam H1. Less intense bands can be seen at a size greater than 20 Kb in the Eco R1 digest , at 4.4 Kb and 1.5 Kb in Hind III, at 2.9 Kb Taq 1, and at greater than 20 Kb in Bam H1. A strong band was obtained at 1.5 Kb in the 6.01 single copy standard track. The results obtained using 5' flanking sequence as a probe are shown in fig. 7. Here several bands were present in each track, the strongest being in the Eco R1 digest at 3.5 Kb and 12 Kb, in Hind III at 0.8 Kb (with feint bands at 4.4 Kb abd 2.7 Kb), in Taq 1 at 3.0 Kb and in Bam H1 at approx 15 Kb and 2.9 Kb. The 6.03 single copy standard gave a strongly hybridising band at 0.8 Kb. The presence of a band at 2.7 Kb in three of the digests was possibly due to contamination of these samples with pUC8 hybridising to traces of itself present in the probe.

Fig.5. Pea Genomic DNA Digests:

Track	1 and 7	10 ug Pea Leaf genomic DNA digested with Eco RI
"	2 and 8	" " " " " " " Hind III
"	3 and 9	" " " " " " " Taq I
"	4 and 10	" " " " " " " Bam H I
"	5	*Leg* D single copy 1 gene equivalent
"	6	Gap
"	11	*Leg* D 5' flanking sequence single copy equivalent

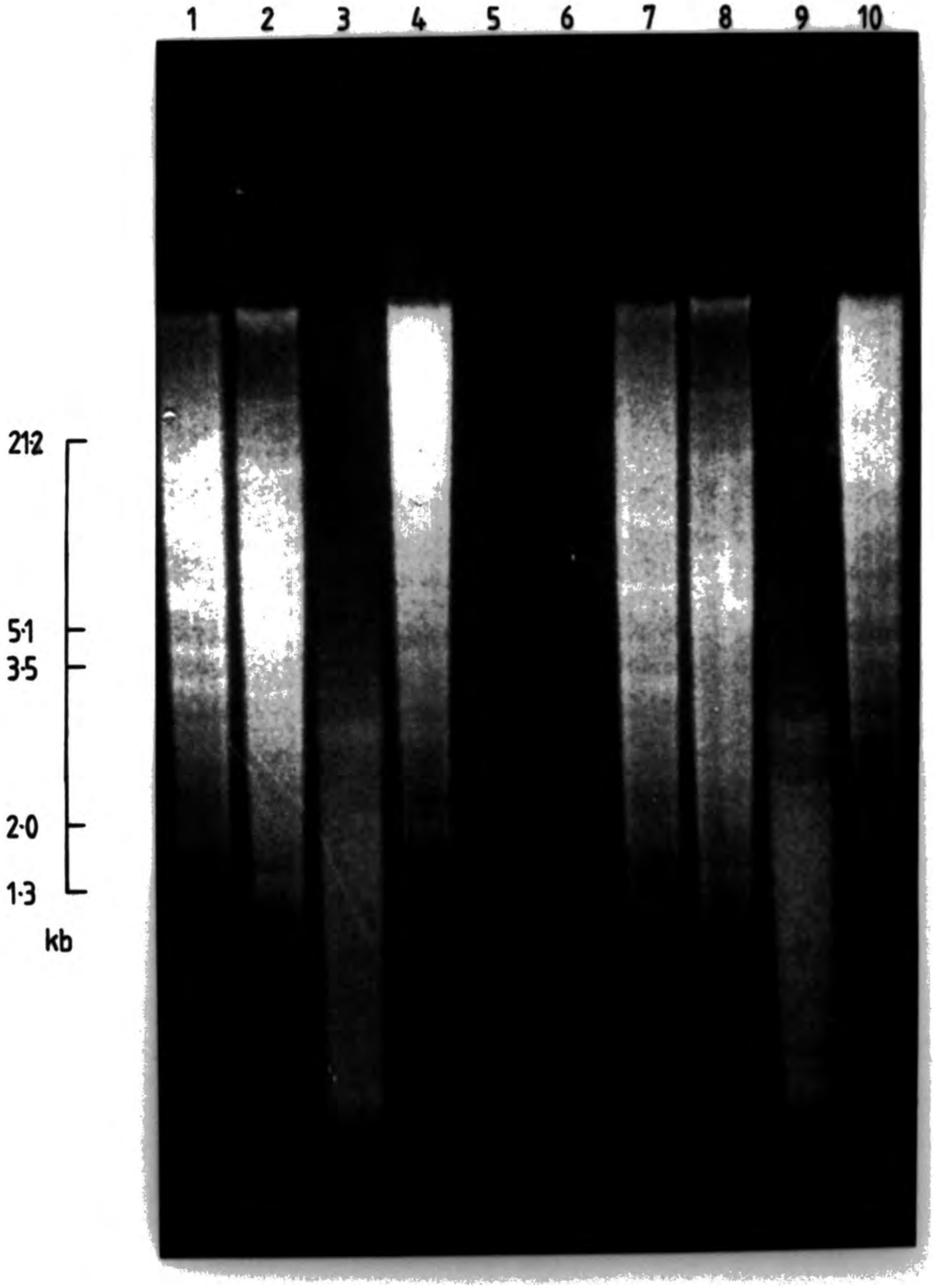

1
2
3
4
5
6
7
8
9
10
21·2
5·1
3·5
2·0
1·3
kb

Fig.6. Hybridisation of *Leg* D (Hind III fragment a, fig.1) to Pea genomic DNA:

Track 1 10 ug Pea Leaf genomic DNA digested with Eco RI
" 2 " " " " " " " Hind III
" 3 " " " " " " " Taq I
" 4 " " " " " " " Bam HI
" 5 *Leg* D single copy gene equivalent

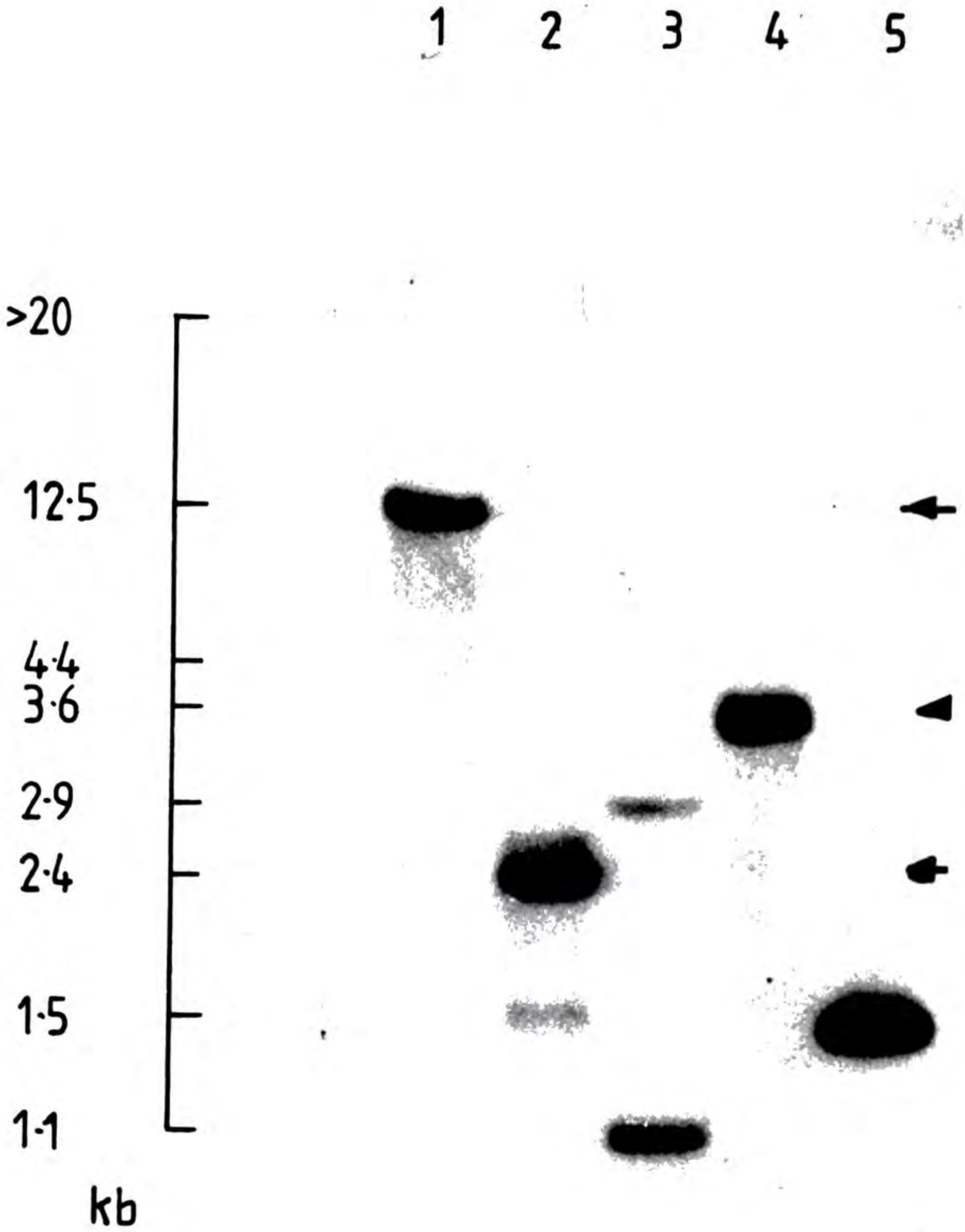
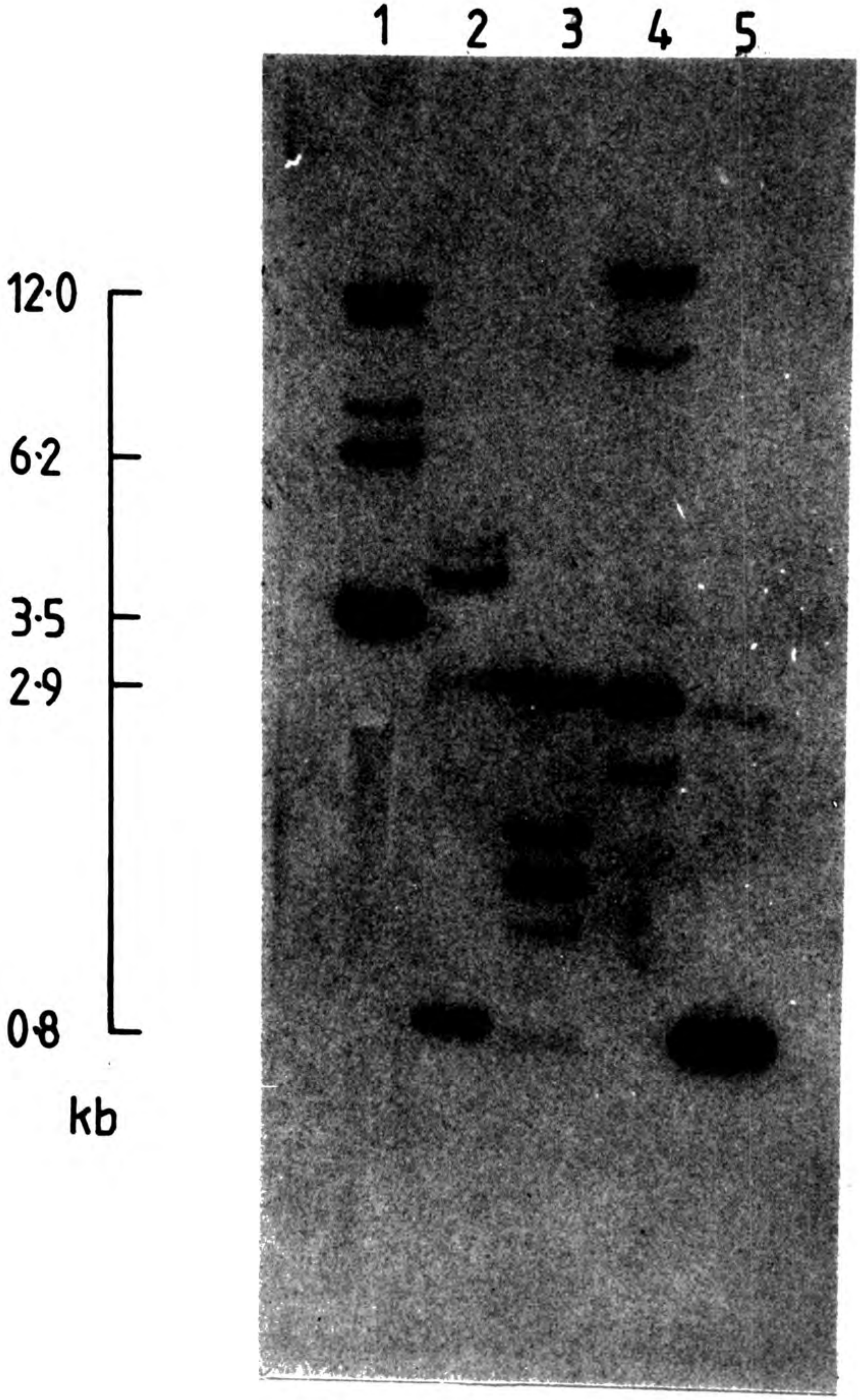

1
2
3
4
5
>20
12·5
4·4
3·6
2·9
2·4
1·5
1·1
kb

Fig.7. Hybridisation of *Leg* D 5' flanking sequence (Hind III fragment b, fig.1) to Pea Genomic DNA:

Track	1	10 ug	Pea	Leaf	genomic	DNA	digested	with	Eco RI
"	2	"	"	"	"	"	"	"	Hind III
"	3	"	"	"	"	"	"	"	Taq I
"	4	"	"	"	"	"	"	"	Bam HI
"	5	*Leg* D 5' flanking sequence single copy equivalent							


1
2
3
4
5
12·0
6·2
3·5
2·9
0·8
kb

3.2 Characterisation and Nucleotide Sequence of Legumin gene K (*Leg* K)

3.2.1. Characterisation of *Leg* K

λ JC-5 is a genomic clone containing a 13.5 Kb fragment of pea genomic DNA produced by partial digestion with Eco RI. Hybridisation of pCD40, a cDNA coding for a 'minor' legumin polypeptide of 65,000 Mr (Domoney and Casey, 1984) to an Eco RI digestion of the genomic clone gave two separate hybridising fragments of 1.9 and 3.5 Kb (Ellis T.H.N. pers. comm.) as indicated in the restriction map of λ JC-5 shown in fig.8.

The 1.9 Kb Eco RI fragment was subcloned into pUC8 and named pJC 5-2. Full characterisation by restriction mapping and nucleotide sequencing confirmed that it contained a gene, Legumin J (*Leg* J), coding for a minor legumin polypeptide (Bown et al., in press). On the strength of this, it was postulated that the 3.5 Kb Eco RI fragment might contain a tandem repeat of this gene, provisionally termed *Leg* K, coding for another minor legumin polypeptide. Hence this 3.5 Kb fragment was subcloned into pUC 8, named pJC 5-11 and restriction mapped using the following restriction enzymes; Alu I, Bal I, Eco RI, Hind III, Pst I, Kpn I, Sau 3A, Rsa I, and Xho I. The resulting map is given in fig. 9, along with a restriction map of *Leg* J, aligned by shared restriction sites, giving the first evidence that the two genes do appear to be tandem repeats of each other.

Fig. 8. Partial restriction map of the pea genomic clone λ JC--5. Bars represent areas corresponding to the coding sequences plus introns of *Leg* J and *Leg* K. Arrows labelled 5.11 and 5.2 represent subclones of *Leg* K and *Leg* J respectively, used for further restriction mapping and sequencing. Restriction sites: B = Bam H1, E = Eco RI, H = Hind III, X = Xho I.

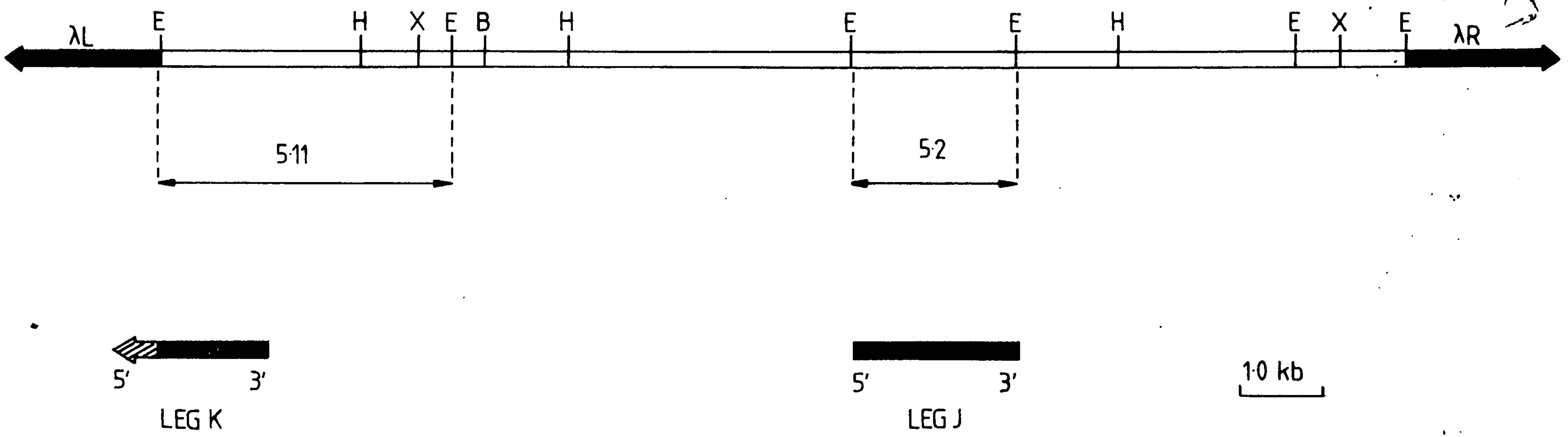
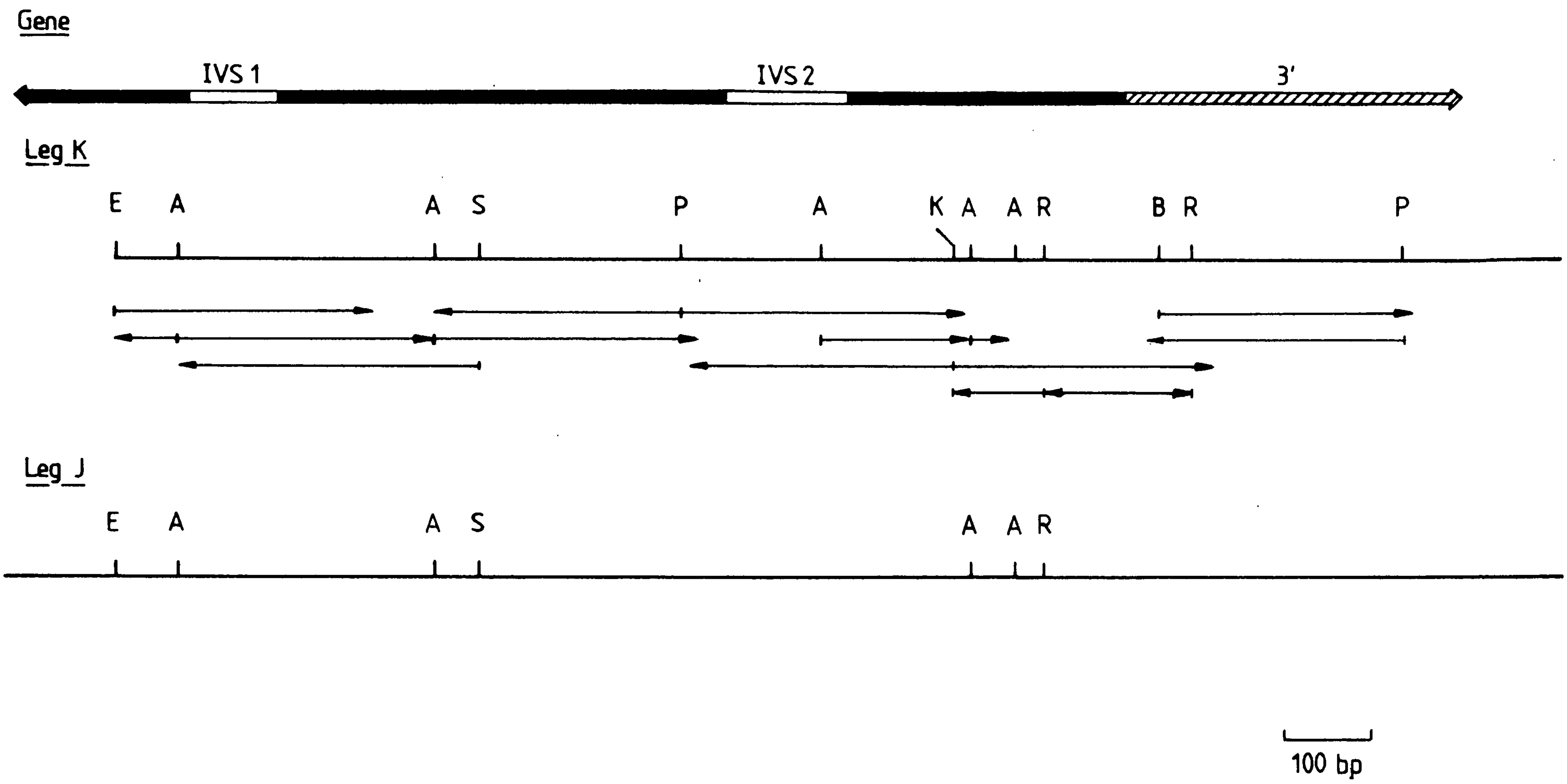

λL
E
H
X
E
B
H
E
E
H
E
X
E
λR
5·11
5·2
5′
3′
LEG K
5′
3′
LEG J
1·0 kb

Fig. 9. Restriction and sequencing maps for *Leg* K, with partial restriction map of *Leg* J showing comparable restriction sites. Restriction sites: A = Alu I, B = Bal I, E = Eco RI, K = Kpn I, P = Pst I, R = Rsa I, S = Sau 3A. Arrows indicate direction and length of sequence obtained from each M13 subclone. Duplicated sequencing runs carried out to check sequences are not shown.
■ = *Leg* K coding sequence, □ = *Leg* K intron sequence,
▨ = *Leg* K 3' flanking sequence.


Gene
IVS 1
IVS 2
3'
Leg K
E
A
A
S
P
A
K
A
A
R
B
R
P
Leg J
E
A
A
S
A
A
R
100 bp

3.2.2. Nucleotide Sequence of *Leg* K

Nucleotide sequence accounting for 1260 bp of coding sequence and 273 bp of 3' flanking sequence was determined by M13 sequencing. Using the following restriction enzymes - Sau 3A, Rsa I, Alu I, Bal I, Kpn I, Pst I fragments from the 3.5Kb Eco RI insert of pJC 5-11 were subcloned into M13 and sequenced by the M13 dideoxy chain termination method (see section 2.9). The nucleotide sequence obtained from *Leg* K is given in fig. 10 where it has been aligned for maximum homology against that of *Leg* J. The sequencing strategy adopted to achieve this result is indicated by the arrows in fig.9, showing subclones used, and the direction of sequence given by them.

Unfortunately, the 5' flanking sequence and part of the 5' coding region were not available for sequencing as they had become detached from the rest of the gene during partial EcoRI digestion of the genomic DNA prior to ligation into the λ vector. Screening of the λ gene library failed to isolate the λ genomic clone containing this missing region of *Leg* K.

Fig.10. Nucleotide sequence of *Leg* J and all that achieved for *Leg* K, aligned for maximum homology. The amino acid sequence encoded by *Leg* J is shown, and those amino acids differing in *Leg* K are indicated above. The N-terminals of the a- and b- subunits of the mature protein are indicated by colons. Transcription start is indicated by over-dots : the base designated by +1 is shown by ^ . Other features are as indicated: 'legumin box' designates the putative 5' enhancer sequence element.

```
LegJ .........(-562).GTTAACACAAGCTAAAATTTATTTGTGCAATCATCATCATGTCATCTTCATCTTCTAATTTGAAATGAAAATTTAGCAAAATACATAACCAGTCAATCTAGAAT -459

LegJ TTACCTAAAAGAGAGACAACTGTATCTATATTATATCAGGGAGTAATACACCAGCAGTACATTTTGAGTGGAGGAGGCCAATTATTAAAGTTTATAAAGTAGTAAAACATGCAAGAGTCG -339

LegJ AATGAAATATATGCTCTAGACAGTAAATTAATAGTTGAGTTAAAGAGATAAATGCATAGAGTCGACGCAGAGAAAAGAACTAGAGAAGTGAAGGGGACCATCCACATATAAGAATACCAA -219

LegJ CAAATATTCATTGTCTCTTTGTGGTATTTGGATATATACTAATTATCAATCTGTGAAGAATGAATGAAGCGGCTACTTGCGCTGCGTCCCACATATGATGTGTATCAATTTAGGACTCCA  -99
     ....................................................................................................................<...

LegJ TAGCCATGCATGCTGAACAATGTCATACACATTCTGTCACACGTGTTCCTATCTCACCCTTCCCCTCTTCCTATAAATCACCACAACACAGCTTCTCCACTTCACCACTTCACTCACCAA   22
     ..."Legumin" BOX.....>................................................<TATA BOX>..................^.....................

LegJ TCTCTCCTTAGTAGTTTATGATCAGAGTCACAATGTCCAAACCTTTTCTATCTTTGCTTTCACTTTCCTTGCTACTCTTTGCAAGCGCATGTTTAGCAACTAGCTCTGAGTTTGACAGAC  142
A.A. ...............................<M  S  K  P  F  L  S  L  L  S  L  S  L  L  L  F  A  S  A  C  L  A :T  S  S  E  F  D  R

LegJ TTAACCAATGCCAGCTAGACAGTATCAATGCATTGGAACCAGACCACCGTGTTGAGTCCGAAGCTGGTCTCACTGAGACATGGAATCCAAATCACCCTGAGCTAAAATGCGCCGGTGTGT  262
A.A. L  N  Q  C  Q  L  D  S  I  N  A  L  E  P  D  H  R  V  E  S  E  A  G  L  T  E  T  W  N  P  N  H  P  E  L  K  C  A  G  V

LegJ CACTTATTAGACGCACCATCGACCCTAATGGACTCCACTTGCCATCTTTCTCCCCCTCTCCACAGTTGATTTTCATCATCCAAGGAAAGGGTGTTCTTGGACTTTCATTTCCTGGCTGTC  382
A.A. S  L  I  R  R  T  I  D  P  N  G  L  H  L  P  S  F  S  P  S  P  Q  L  I  F  I  I  Q  G  K  G  V  L  G  L  S  F  P  G  C

LegJ CTGAGACTTATGAAGAGCCTCGTTCATCACAATCTAGACAAGAATCCAGGCAGCAACAAGGTGACAGTCACCAGAAGGTTCGTCGATTCAGAAAAGGTGATATCATTGCCATTCCATCGG  502
A.A. P  E  T  Y  E  E  P  R  S  S  Q  S  R  Q  E  S  R  Q  Q  Q  G  D  S  H  Q  K  V  R  R  F  R  K  G  D  I  I  A  I  P  S

LegK GAATTCCTTATTGGACATATAACCATGGCGATGAACCTCTTGTTGCCATTAGCCTTCTTGACACTTCCAACATTGCAAACCAGCTCGATTCAACCCCAAGAGTAAGTGATAGTGTATCCA
LegJ GAATTCCTTATTGGACATATAACCATGGCGATGAACCTCTTGTTGCCATTAGTCTTCTTGACACTTCCAACATTGCAAACCAGCTCGATTCAACCCCAAGAGTAAGTAATAGTGTATCCA  622
A.A. G  I  P  Y  W  T  Y  N  H  G  D  E  P  L  V  A  I  S  L  L  D  T  S  N  I  A  N  Q  L  D  S  T  P  R <.................

LegK TTCAT--------------------------ACAGTATGCTCTTTCGATTATAACTT-AAAAGTTTCTAAT------------------------------GTAAATATGTGTATGCAGG
LegJ TACATTACATTATCTCTTATAAATTGTTCATACAGCATGCTCATTCGATTATAACTTTAAAAGTTTCTAATGTATAATTTGTTATACTAAATCAATCACACGTAAATATGTGTATGCAGG  742
A.A. ............................................ Intron-1 .................................................................>

LegK TATTTTACCTTGGTGGAAACCCAGAAACAGAGTTCCCCGAAACACAGGAGGAACAACAAGGAAGGCATCGGCAAAAGCATAGTTACCCTGTTGGACGTAGGAGTGGACATCACCAACAAG
LegJ TATTTTACCTTGGTGGGAACCCAGAAACAGAGTTCCCCGAAACACAGGAGGAACAACAAGGAAGGCATCGGCAAAAGCATAGTTACCCTGTTGGACGTAGGAGTGGACATCACCAACAAG  862
A.A. V  F  Y  L  G  G  N  P  E  T  E  F  P  E  T  Q  E  E  Q  Q  G  R  H  R  Q  K  H  S  Y  P  V  G  R  R  S  G  H  H  Q  Q

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

A.A.                                                 N                 E                                               E
LegK ACGAAAGGAGTCAAATTGTGCGAGTTGAGGGAGGTCTCCGCATTATCAACCCCAAGGGGAAGGAAGAAGAAGAAGAAAAAGAACAGAGTCATTCTCACTCTCACAGAGAGGAGGAGGAAG
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
```

```
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 :G  L  E  E  T  I  C  S  A  K  I  R  E  N  I  A  D  A  A  R

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

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

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

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 TATGATAATGCAATACAATAACAAGATGGCCATCTTGTCTTGAATAATAATGATAAAATAAATTTTGAATGAACTGTAGTAAAAAAATTTCACTTCCTATAATAAAACAATAAAGTATCGC 1930
     ....................................................:..<PolyA+>...................................<PolyA+><PolyA+>......

LegK CTTAAAATCCCAATCTTAATCTAAATTTGTATGCATCTATA-AGGGGCGAATAACACTAGTTTTGTT-CACCTTGCAATTGCCATAATAAAATG-CATA-CACTTTTTACTATTGCTTAT
LegJ CCTACTACCCTGATCTTACTCTGAATTTGTATGCATGTAAAGAGGGGTGAATAACAATTGGTTTTGTACACCTTCCAATTGCCATAATAAAATGGCATATCACTTTTTAAAAAATTCTCC 2050
     ....................................................................................<PolyA+>............................

LegK ATGTTTGTTTGAATCATATAAAAAACACAACTACAAATCTGCATTTTTCTTCGGCATTTGATTATATATCTGCAG.............................................
LegJ TTCATTATCTTCTTCATTAGTACTAATGATGAATTGTCTCAATAATAATATCAGCTTTTTGAATACAGCAACGAGACAGCAAACTTTAACAATCACAATTATAAGGTAGTGTTATTAATT 2170

LegJ GTTTTCTAAAAATCTTTTCTGTGTGACGGAAGGAATCATTCATTTCCTTATCAATATAATTGTCAAACTATAAAGGATAAAAATAAATACTTGAGAGGCAAGAGGTAAGAAGATGTGGTT 2290

LegJ AATCCCACTACTCGCTAGAGAATGAACACACTTGTCATGTTGATAAAAATGAGAACCGAAAAATCTACAAATAAATACTAGTTGATTAAATAAAGAATTTGTATTGTTGATAAATACTGA 2410

LegJ ATTC....(2414)
```

3.2.3. Hybridisation of *Leg* K to messenger RNA

A 3' flanking sequence probe specific to *Leg* K was prepared by isolating and labelling the 2.3 Kb Bal 1/Eco RI fragment 3' to the Bal I site in pJC 5-11 (see fig.9). This was then hybridised to a Northern blot of total pea cotyledon RNA, resulting in the detection of two bands. One was an intense band at an indicated size of approx 2,300 bases, and the other, feinter band was of approx 4,200 bases. (see fig.11).

Pictured alongside the result shown in fig.11 are results produced by D. Bown, where probes corresponding to *Leg* J coding sequence and 3' flanking sequence (specific only to *Leg* J) were hybridised independently to Northern blots of total RNA extracted from cotyledons in varying degrees of development. The coding sequence probe hybridised to a heterogeneous RNA population in the approx. size range of 2100 - 2400 bases, while the 3' flanking sequence probe hybridised to a single band of approx. 2100 bases. Intensity of hybridisation increases as the stage of development of the cotyledons from which the RNA was extracted increases.

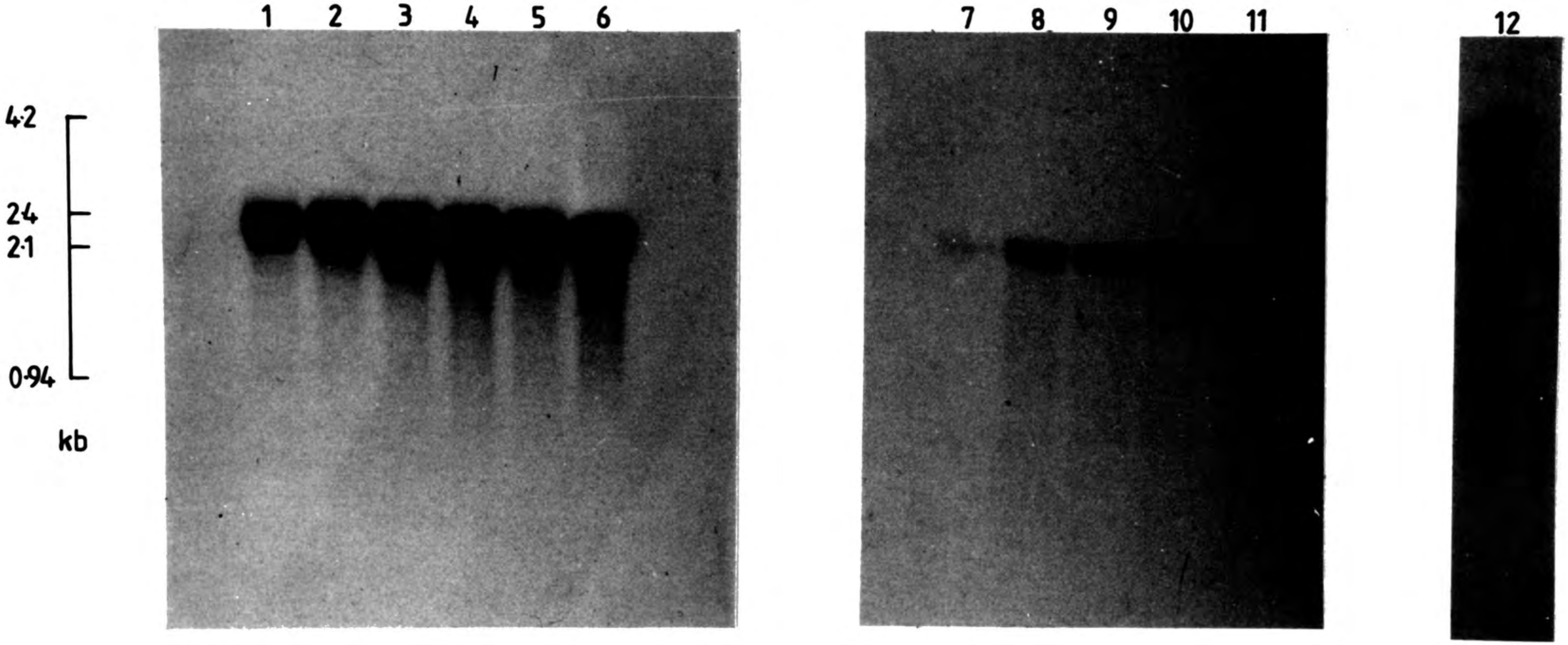
Fig.11. Hybridisation of *Leg* J and *Leg* K sequences to total RNA prepared from developing pea cotyledons ('Northern' blots).

Tracks 1 - 6 , *Leg* J coding sequence probe hybridised to total RNA from cotyledons
1 - 8-9 d.a.f. (days after flowering)
2 - 10 "
3 - 12 "
4 - 14 "
5 - 16 "
6 - 18 "

Tracks 7 - 11, 3' flanking sequence probe from *Leg* J hybridised to total RNA from cotyledons
7 - 11 d.a.f.
8 - 12 "
9 - 14 "
10 - 18 "
11 - 20 "

Track 12, 3' flanking sequence probe from *Leg* K hybridised to total RNA from pea cotyledons 18 d.a.f.

Cotyledon expansion takes place over the period 7.8 d.a.f. under the conditions used (Gatehouse et al., 1982)


1
2
3
4
5
6
7
8
9
10
11
12
4·2
2·4
2·1
0·94
kb

3.3 Characterisation and Complete Nucleotide Sequence of Vicilin Gene B (*Vic* B), coding for a 50,000 Mr Vicilin Polypeptide

3.3.1. Characterisation of *Vic* B

λ JC-1 is a genomic clone containing a 13.8 Kb fragment, and was isolated from a λ gene bank containing fragments resulting from an Eco RI partial digest on pea leaf genomic DNA (Ellis et al., 1986). A cDNA, pCD48, coding for a 50,000 Mr vicilin polypeptide (Domoney and Casey, 1985) was used in the screening process. Eco RI digestion of λ JC-1 produced 4 fragments, one of which, at 4.7 Kb, proved homologous to pCD48 in further hybridisation studies. (see fig.12). This 4.7 Kb Eco RI fragment has been sub-cloned into pUC9 (Ellis et al., 1986) and termed pJC 1-16.

In order to locate the gene on the 4.7 Kb Eco RI fragment and also to determine its orientation therein on both the genomic clone λ JC-1 and the pUC9 subclone pJC1-16, a cDNA, p DUB9, encoding a 50,000 Mr vicilin polypeptide, (Delauney, 1984) was digested with Bgl II to produce 0.36 Kb 5' and 1.1 Kb 3' coding sequence specific probes. These were then hybridised separately to two identical southern blots of pJC 1-16 double digests consisting of one enzyme - either Sst I (one site in the pJC-16 insert - see fig.12) or Bam HI (no sites in the insert but one site in pUC9) to linearise the plasmid in conjunction with one of a range of 6 bp recognition site enzymes. An example of this is shown in fig. 13, showing the band sizes obtained from an agarose gel containing a set of digests on pJC-16, and from the resulting autoradiograph, those bands which hybridised to the 5' coding sequence probe. After the gene had been located and orientated on the fragment in this way, restriction mapping was completed to give the final map showing major restriction sites, the approximate location of the gene, and the orientation within pJC 1-16 (see fig 14). According to these results, the gene appeared to be contained in its entirety within the 4.7 Kb Eco RI fragment.

Fig. 12. Partial restriction map of the pea genomic clone λ JC-1. The bar represents the area corresponding to the coding sequence plus introns of *Vic* B. The arrow labelled 1-16 represents the subclone used for further restriction mapping and sequencing of *Vic* B. Restriction sites : E = Eco RI, H = Hind III, S = Sal I, T = Sst I.

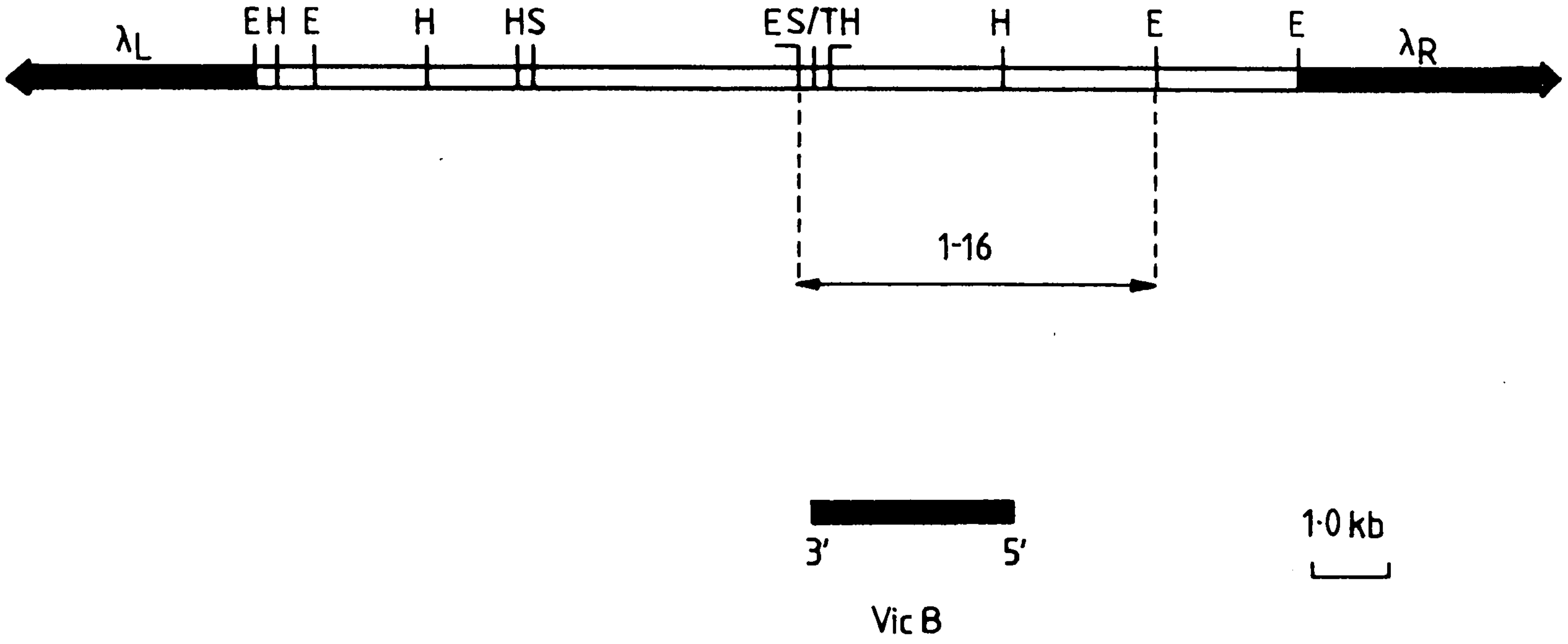

λL
E H E
H
HS
E S/T H
H
E
E
λR
1-16
3'
5'
Vic B
1·0 kb

Fig. 13. Example restriction digestion of *Vic* B subclone pJC 1-16(A) plus hybridisation of the 5' Bgl II section of pDUB9 (see 3.3.1.) to a Southern' blot of these digests (B)

A: Track	1	λ DNA	digested	with	Eco RI	+ Hind III
	2	pJC 1-16	"	"	Hpa I	+ Sst I
	3	"	"	"	Pvu II	+ Bam HI
	4	"	"	"	Xho I	+ Sst I
	5	"	"	"	Bal I	+ Sst I
	6	"	"	"	Bcl I	+ Bam HI
	7	"	"	"	Stu I	+ "
	8	"	"	"	Eco RV	+ "
	9	"	"	"	Acc I	+ "
	10	"	"	"	Hind III	+ "

B: Track	1	pJC 1-16	Digested	with	Hpa I	+ Sst I
	2	"	"	"	Pvu II	+ Bam HI
	3	"	"	"	Xho I	+ Sst I
	4	"	"	"	Bal I	+ Sst I
	5	"	"	"	Bcl I	+ Bam HI
	6	"	"	"	Stu I	+ "
	7	"	"	"	Eco RV	+ "
	8	"	"	"	Acc I	+ "
	9	"	"	"	Hind III	+ "

Restricted fragment sizes (Kb)+ hybridising bands (indicated with *) pJC 1-16 digested with (fragments in brackets were of doubtful visibility)

Hpa I + SstI	3.15	2.5*	1.2	0.52		
Pvu II + Bam HI	4.8*	2.3	(0.3)			
Xho I + Sst I	3.7*	2.8*	1.2	0.92x2	0.74	0.52
Bal I + Sst I	5.0	2.3*	0.8*			
Bcl I + Bam HI		no Bcl I sites				
Stu I + "		no Stu I sites				
EcoRV + "		?				
AccI + "	4.5	1.8*	0.7	0.4		
Hind III+ "	4.9	1.9*	0.5			

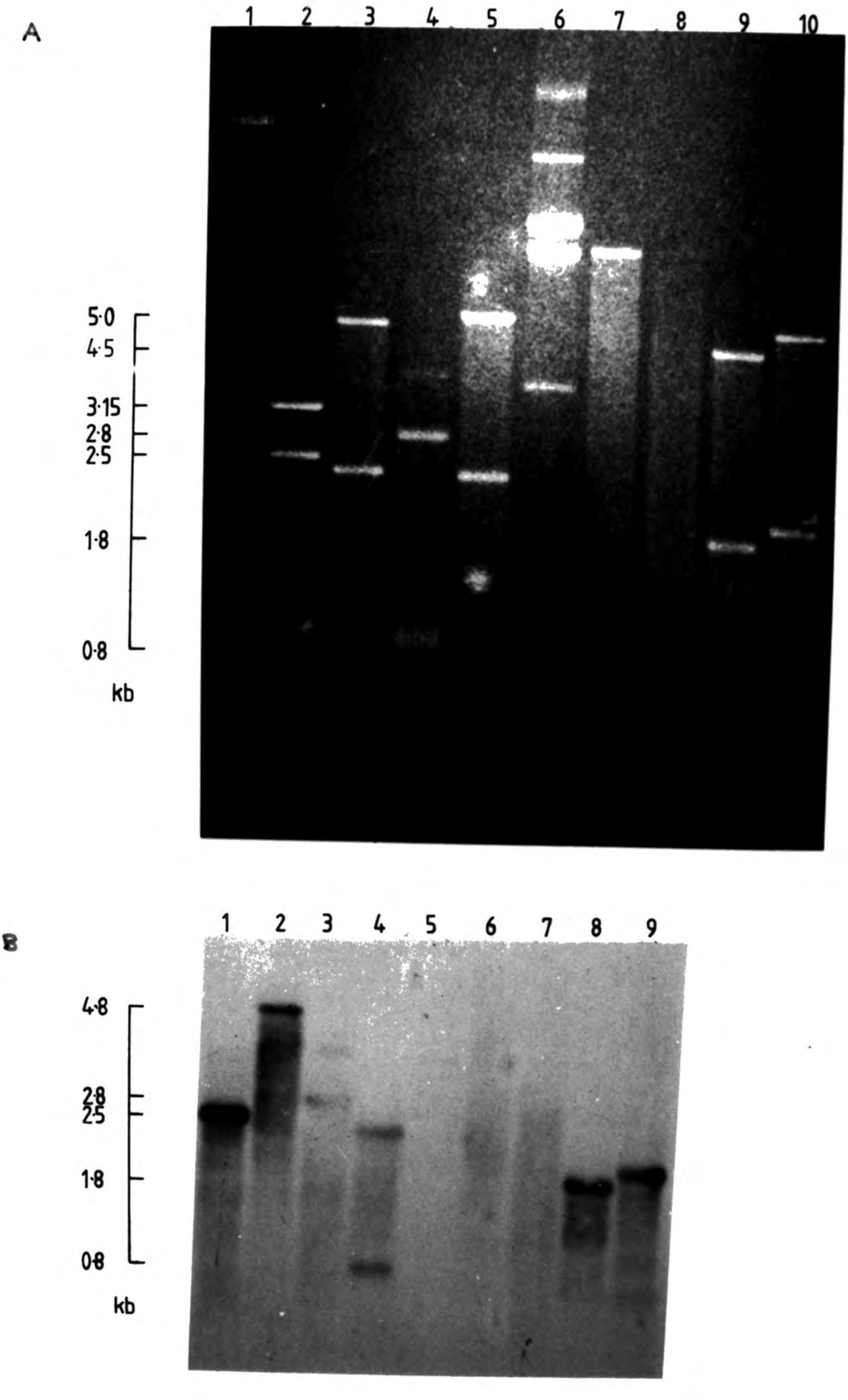
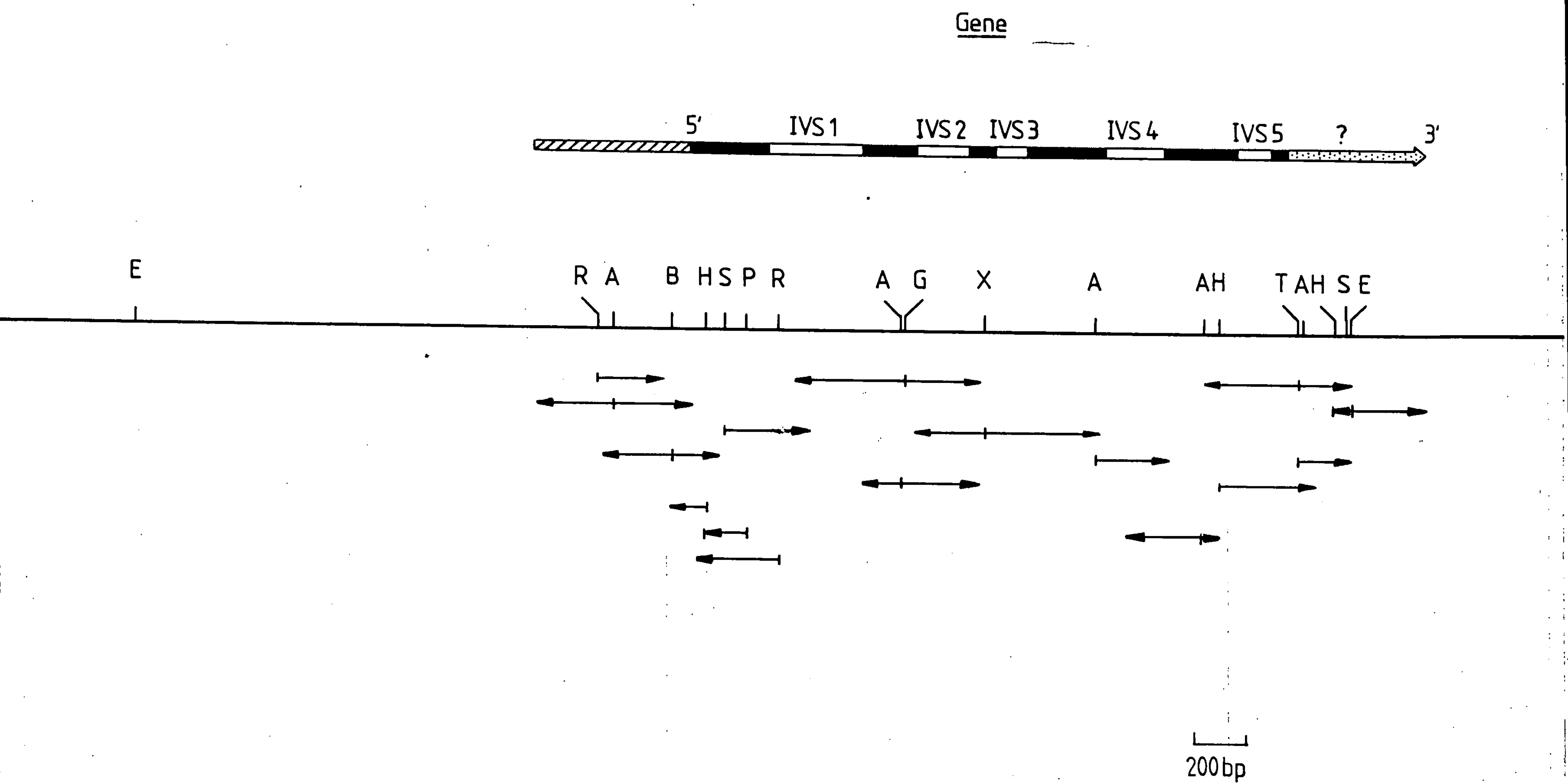

A
1 2 3 4 5 6 7 8 9 10
5·0
4·5
3·15
2·8
2·5
1·8
0·8
kb
B
1 2 3 4 5 6 7 8 9
4·8
2·8
2·5
1·8
0·8
kb

Fig. 14. Restriction and sequencing maps of *Vic* B coding and flanking sequence. Restriction sites : A = Alu I, B = Bal I, E = Eco RI, G = Bgl II, H = Hind III, P = Pst I, R = Rsa I, S = Sau 3A, T = Sst I. Arrows indicate direction and length of sequence obtained from each M13 subclone. Duplicated sequencing runs carried out to check sequences are not shown. ■ = *Vic* B coding sequence, □ = *Vic* B intron sequence, ▨ = *Vic* B 5' flanking sequence, ▒ = 3' sequence of unknown origin.


Gene
5'
IVS 1
IVS 2
IVS 3
IVS 4
IVS 5
?
3'
E
R A
B
H S P
R
A G
X
A
A H
T A H S E
200 bp

3.3.2 Nucleotide Sequence of *Vic* B

The complete nucleotide sequence of *Vic* B is given in fig. 15, and includes 653 bp of 5' flanking sequence, and all the coding sequence up to 35 bp into Exon 6, at which point it appears to be interrupted by sequence of unknown origin. Sequencing strategy for this is shown in fig. 14, again by a series of arrows indicating direction of sequence obtained from a given type of subclone. As can be seen from this, the bulk of sequence information was gleaned from Sau 3A, Rsa 1, and Alu 1 fragments shotgun cloned into M13 (see section 2.9.1.1.). Remaining gaps in the sequence were then filled in by digestion of the pJC 1-16 plasmid with Eco R1/Sst I, Eco RI/Hind III, Eco RI/Bal I, Eco RI/Xba I, Hind III/Pst I, Hind III/Bal I, Hind III/Xba I and Hind III/Bgl II. Fragments from these digestions were then either shotgunned or cloned by the direct method (see section 2.9.1.2.) into appropriately cut M13.

3.3.3 Partial Sequencing of pJC 1-12

In an attempt to locate the missing 3' end of *Vic* B, the pUC 9 subclone pJC 1-12 (Ellis et al., 1986), containing the 6.4 Kb Eco RI fragment immediately adjacent to the 3' end of the 4.7 Kb Eco RI fragment in λ JC-1 (see fig. 12), was digested with Eco RI/Bam HI. The 1.1 Kb fragment from this digest was then subcloned into the appropriate M13 vector to allow 200 bp of sequence to be read out from the Eco RI site at the 3' end of *Vic* B, thus extending sequence at the interrupted end of the gene in order to see if coding sequence resumed in this region (see fig. 14 for the direction of sequence obtained from this subclone). As shown by the sequence given in fig. 15, coding sequence did not resume within this extra 200 bp stretch.

3.3.4 Hybridisation of pDUB9 to pJC 1-12

A Southern blot of pJC 1-12 digested with Eco RI was probed with the homologous Vicilin cDNA pDUB9 (see section 3.3.1). No hybridising bands were detected, (result not presented), indicating that the missing 3' end of *Vic* B was not present within the 6.4Kb Eco RI fragment in pJC1-12.

Fig.15 Complete nucleotide sequence of *Vic* B and its flanking regions. The amino acid sequence is given below each line. Major features are as indicated.

```
                                   ############################################
                                   #                                          #
                                   # NUCLEOTIDE SEQUENCE OF VICILIN GENE vic B #
                                   #                                          #
                                   ############################################


VicB ...............................................................TATATATTATATTTTTCTTTTTAATATATAAATAAAGTATAGTATATGTAAAGTAAA

VicB CGGATAAATAATAGATAAATAATTAAATGACATATATGTACAATTACATTTTTATATATAAATTGACATATATATGTACAATTACATTTTTATATATTAAGTATAGTATAGTATATATAA

VicB AGTAGACGGATAAATAATGATAGATAATTAAATGACGTATATGTACAATTACATTTTTCACATGACAAGTACAAACATATGCACTTCTAAGTGCAAGTTTATGGAGTTATTTGCATGTCT

VicB TAGAGCTGGAGCTTGAGTTGTAGGATACAACACTTGTTAAAATTCTCTAGTCAATTCATTAATTCATATACACATGGCCGAAGACAATAATAAAGCATCCTCCTTTTCCATAAGAATGTC

VicB CAAATTCATCAAATTCAAACAAAACTCCACCACCCAAGTAATGTTCTTTTCATTTTGCCACTTCAATTTTGTACATTTTAACACACGTCCATATGCATGGCACAACATGGCCAACTGTTG

VicB GTGCATGTTAATTATATAGTTTTATTTTTTATATCTATAAATACACTCATCTCACTGTACTTTATTCATCCAGAGCGACCAAAGTGAGATATTAGTTTCAATCAACAGGCTGCTACTACA
     ............................<TATA BOX>..................................................................................

VicB ATGAAAGCTTCATTTCCACTTTTGATGCTAATGGGAATCTCTTTCCTAGCATCAGTGTGTGTTTCTTCTAGGTCTGATCCTCAAAATCCTTTTATCTTCAAGTCTAACAAGTTTCAAACT  120
A.A. <M  K  A  S  F  P  L  L  M  L  M  G  I  S  F  L  A  S  V  C  V  S  S :R  S  D  P  Q  N  P  F  I  F  K  S  N  K  F  Q  T

VicB CTTTTTGAGAATGAAAATGGGCACATTCGACTTCTGCAGAAATTTGACCAACGTTCTAAAATTTTCGAGAATCTACAAAACTACCGTCTTTTGGAATATAAGTCCAAACCTCACACAATA  240
A.A.  L  F  E  N  E  N  G  H  I  R  L  L  Q  K  F  D  Q  R  S  K  I  F  E  N  L  Q  N  Y  R  L  L  E  Y  K  S  K  P  H  T  I

VicB TTTCTTCCACAGCACACCGATGCCGATTACATCCTTGTTGTACTCAGTGGTAATTTATTATTTATCTAAGTTATTATTTATTCTACATCTCTCTATGAGCTTCATTCAAATGCGCGGTAT  360
A.A.  F  L  P  Q  H  T  D  A  D  Y  I  L  V  V  L  S <-----------------------------------------------------------------------

VicB TTATTATTTGTGAGGAGCTGCGGTAGTAACACTCTCTTTCAACACTCTCAATCACTCACTTTTTATGGTTGAAACATGTGGCCTCGCTTGGAAATAGGCCCACATAAAGTGGTAAGAGCA  480
A.A. ---------------------------------------------------------------------------------------------------------------IVS-1----

VicB CACATATTTCAACAAATGAAAGAGTGTGTGTTTGAGAGAGTGTTGAAAAGAGAGTATTATTAGCATTTCTCTTATTACTTTACATATTTGTTTTGTAAGGATAAATTAAATTCAGTTATA  600
A.A. ------------------------------------------------------------------------------------------------------------------------

VicB AAATCTAATGTCATTGTAATTTCCAGGAAAAGCTATACTCACAGTGTTGAATCCCGATGATAGAAACTCCTTCAACCTTGAGCGCGGAGATACGATAAAACTTCCTGCTGGCACAATTGC  720
A.A. -------------------------->G  K  A  I  L  T  V  L  N  P  D  D  R  N  S  F  N  L  E  R  G  D  T  I  K  L  P  A  G  T  I  A

VicB TTATTTGGTTAACAGAGATGACAACGAGGAGCTTAGAGTATTAGATCTCGCCATTGCCGTAAATAGACCTGGCCAACTTCAGGTAATATAACCAATGTTTATCTATTCTCATATCAAATA  840
A.A.  Y  L  V  N  R  D  D  N  E  E  L  R  V  L  D  L  A  I  A  V  N  R  P  G  Q  L  Q <---------------------------------------------------

VicB TGCTATGCATTCTAATGTACAAACAAATGTTAGGGGCCTCTACCATAACATCACAACAAAAATTGCGCCTGTACATATTTTTCTGTAATATTTTCCTAATATTTTCTTTATTTTTTTTGT  960
A.A. ----------------------------------IVS-2---------------------------------------------------------------------------------

VicB CCTTTTTCAACAGTCTTTCTTATTGTCTGGAAATCAAAACCAACAAAACTACTTATCTGGGTTCAGTAAGAACATTCTAGAGGCTTCCTTCAATGTAAGCATAACACACAATTTTTTTTT  1080
A.A. ----------> S  F  L  L  S  G  N  Q  N  Q  Q  N  Y  L  S  G  F  S  K  N  I  L  E  A  S  F  N <-------------------------

VicB CATTTATGTATGTATTAGTTTGGTATTGTATATGTTAATACTCACTTTGTCAATGTATGTACTGTAAAAAAATATAGACTGATTATGAAGAGATAGAAAAGGTTCTTTTAGAAGAGCATG  1200
A.A. --------------IVS-3---------------------------------------------------------> T  D  Y  E  E  I  E  K  V  L  L  E  E  H

VicB AGAAAGAGACACAACACAGAAGAAGCCTTAAGGATAAGAGGCAGCAAAGTCAAGAAGAGAATGTAATAGTAAAATTATCAAGGGGACAAATTGAGGAATTGAGTAAAAATGCAAAGTCTA  1320
A.A. 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  S  R  G  Q  I  E  E  L  S  K  N  A  K  S

VicB CCTCCAAAAAAAGGTGTTTCCTCTGAATCTGAACCATTCAACTTGAGAAGTCGCGGTCCTATCTATTCCAACGAGTTTGGAAAATTCTTTGAAATCACCCCAGAGAAAAATCCACAGCTTC  1440
A.A. T  S  K  K  G  V  S  S  E  S  E  P  F  N  L  R  S  R  G  P  I  Y  S  N  E  F  G  K  F  F  E  I  T  P  E  K  N  P  Q  L

VicB AAGACTTGGATATATTTGTCAATTCTGTAGAGATTAAGGAGGTATGATAAAATTATTTTATAATATAGGAAATTCACCAAATTACACAATGAGATTTCACTTGATCAAATTACAATTGTT  1560
A.A. Q  D  L  D  I  F  V  N  S  V  E  I  K  E <-----------------------------------------------------------------------------------

VicB CTAAATGATTTGATTTTTGTCCTTTGAAGTTATAATGTCAAACTTTTGTTACTAACTTGACATCTCATACACAACAAGTTTTACATACTCAATAACATGTTTTATTTATAGAACATATAT  1680
A.A. ------------------------------IVS-4-------------------------------------------------------------------------------------
```

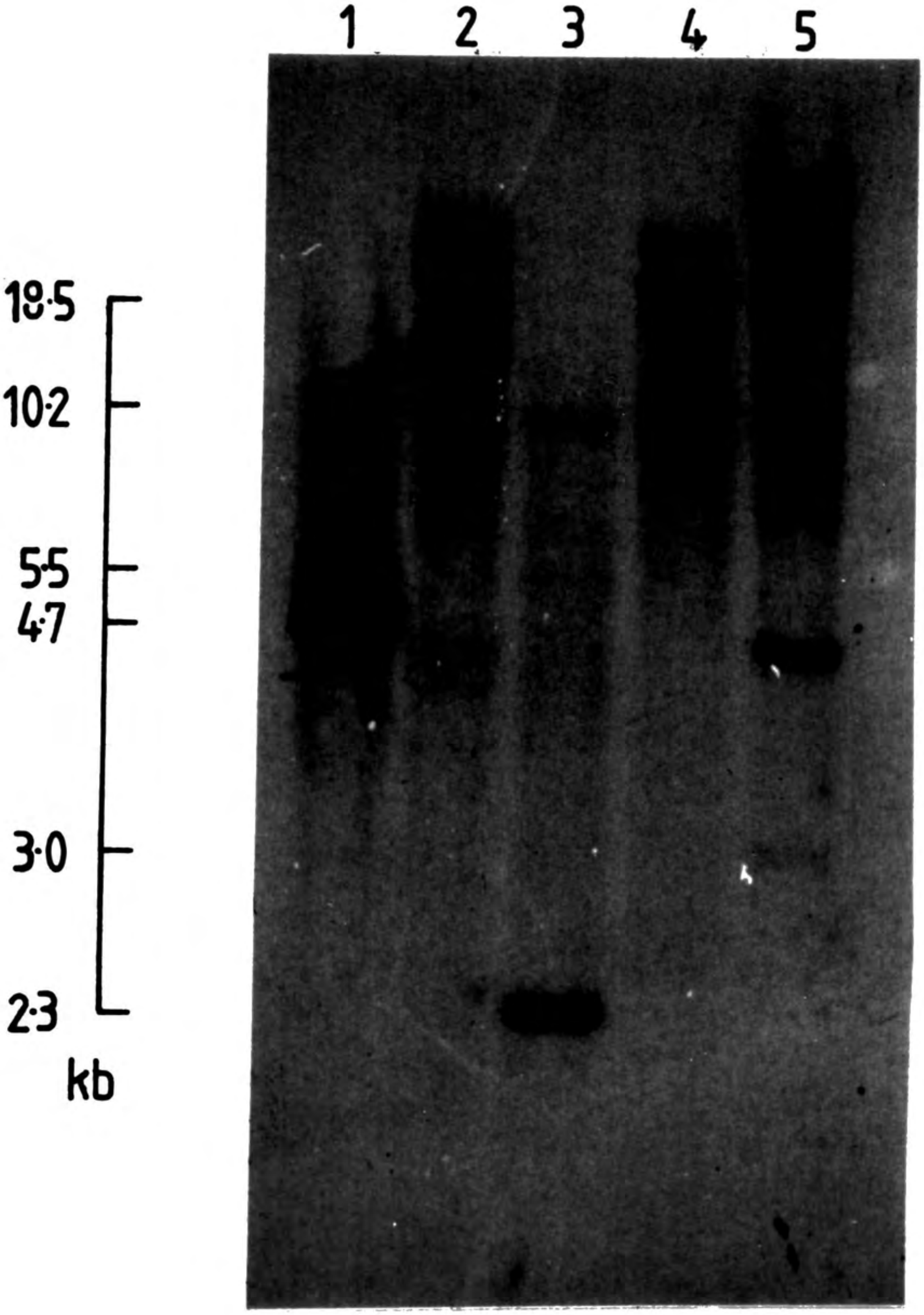
```
VicB CTAATGTATTTATTTAATTATTCTTTCAAATTAAATATTAGGGATCTTTATTGTTGCCACACTACAATTCAAGGGCCATAGTAATAGTAACAGTTAACGAAGGAAAAGGAGATTTTGAAC  1800
A.A. ---------------------------------------->G  S  L  L  L  P  H  Y  N  S  R  A  I  V  I  V  T  V  N  E  G  K  G  D  F  E

VicB TTGTGGGTCAAAGAAATGAAAACCAACAAGAGCAGAGAAAAGAAGATGACGAGGAAGAGGAACAAGGAGAAGAGGAGATAAATAAACAAGTGCAAAATTACAAAGCTAAATTATCTTCAG  1920
A.A. L  V  G  Q  R  N  E  N  Q  Q  E  Q  R  K  E  D  D  E  E  E  E  Q  G  E  E  E  I  N  K  Q  V  Q  N  Y  K  A  K  L  S  S

VicB GAGATGTTTTTGTGATTCCAGCAGGCCATCCAGTTGCCGTAAAAGCAACCTCAAATCTTGATTTGCTTGGGTTTGGTATTAATGCTGAGAACAATCAGAGGAACTTTCTTGCAGGTATAT  2040
A.A. G  D  V  F  V  I  P  A  G  H  P  V  A  V  K  A  T  S  N  L  D  L  L  G  F  G  I  N  A  E  N  N  Q  R  N  F  L  A  <-----

VicB TATATTATCACCCAGTCTCTGTCACTATTTATTCATTTTAAGTGTGTATTTTAAAAGTCGACTTCTATTTAAATCAAGGGGAAAATATTAAGATATGCTTATTATTTTGGTGATTAAAAA  2160
A.A. ------------------------------------------------------------IVS-5------------------------------------------------------

VicB TTTGAAGGCGATGAGGATAATGTGATTAGTCAGATACACCGAGTATATTCTTGGAGCTGAAACTATCCGTTGCATGTTAGAGCTCTCTGAAACCAAGATTTTTAAGATTCCCTAAGCTAA  2280
A.A. ------>G  D  E  D  N  V  I  S  Q  I  H  R <----------------------------------------IVS-6?----------------------------

VicB CAACAGCCTCCTCTTTATAATCACCAATATACAAGATAACTCCATGCCGGTTCCAATGCAGGTTCCTCTTCTTCAGGTTCATATGCTTCCGCAAGCTTACCATCATGATTCGAACACGGT  2400
A.A. ------------------------------------------------------------------------------------------------------------------------

VicB AGAGTTCGAAACAATTTTGCCGCTGCCATCTTAGTGATCTCCTGCATTGTCGCTTCACTGAACTTTGAATTCGCTGAAATATAATCAACAAGTTCAAGCAAAGTTTGCCTCTCTCCAATG  2520
A.A. ------------------------------------------------------------------------------------------------------------------------

VicB ATACGATCCCTGTTCCTTCTACTACCGGCAAGAACTTTTGGTATTTTCCTTTAACTTGTTCTGTTTCTGCAGGTTCGGCGGTTTTATGCATAATGGCTTTTGGTCGTAGTCAAAACCGTG  2640
A.A. ------------------------------------------------------------------------------------------------------------------------

VicB CTCTCACACCAAGCTTCCTAACAACATCATCAAACTGTGGCAAC  2684
A.A. --------------------------------------------
```

3.3.5 Hybridisation of *Vic* B to Genomic DNA

The 1.95 Kb Hind III fragment from pJC 1-16, containing nearly all the *Vic* B coding sequence (see fig. 14) was hybridised to a Southern blot of pea leaf genomic DNA restricted with 5 different 6 bp recognition site restriction enzymes. The resulting autoradiograph is pictured in fig. 16, with the most intense hybridisation occurring at 4.7 Kb in the Eco RI digest, and at 2.3 Kb with Hind III. Feint bands can be observed at 5.5 Kb in the Eco RI track, at approx 22, and 11.5 Kb in the Bam HI, at 9.4 Kb in Hind III, at approx. 24, 18.5 and 10.2 Kb in Eco RV and at 4.5 and 3.0 Kb in the Bgl II track.

Fig. 16 Hybridisation of *Vic* B coding sequence to genomic DNA

Track 1 10 ug Pea leaf genomic DNA digested with Eco RI
2 " " " " " " " Bam HI
3 " " " " " " " Hind III
4 " " " " " " " Eco RV
5 " " " " " " " Bgl II

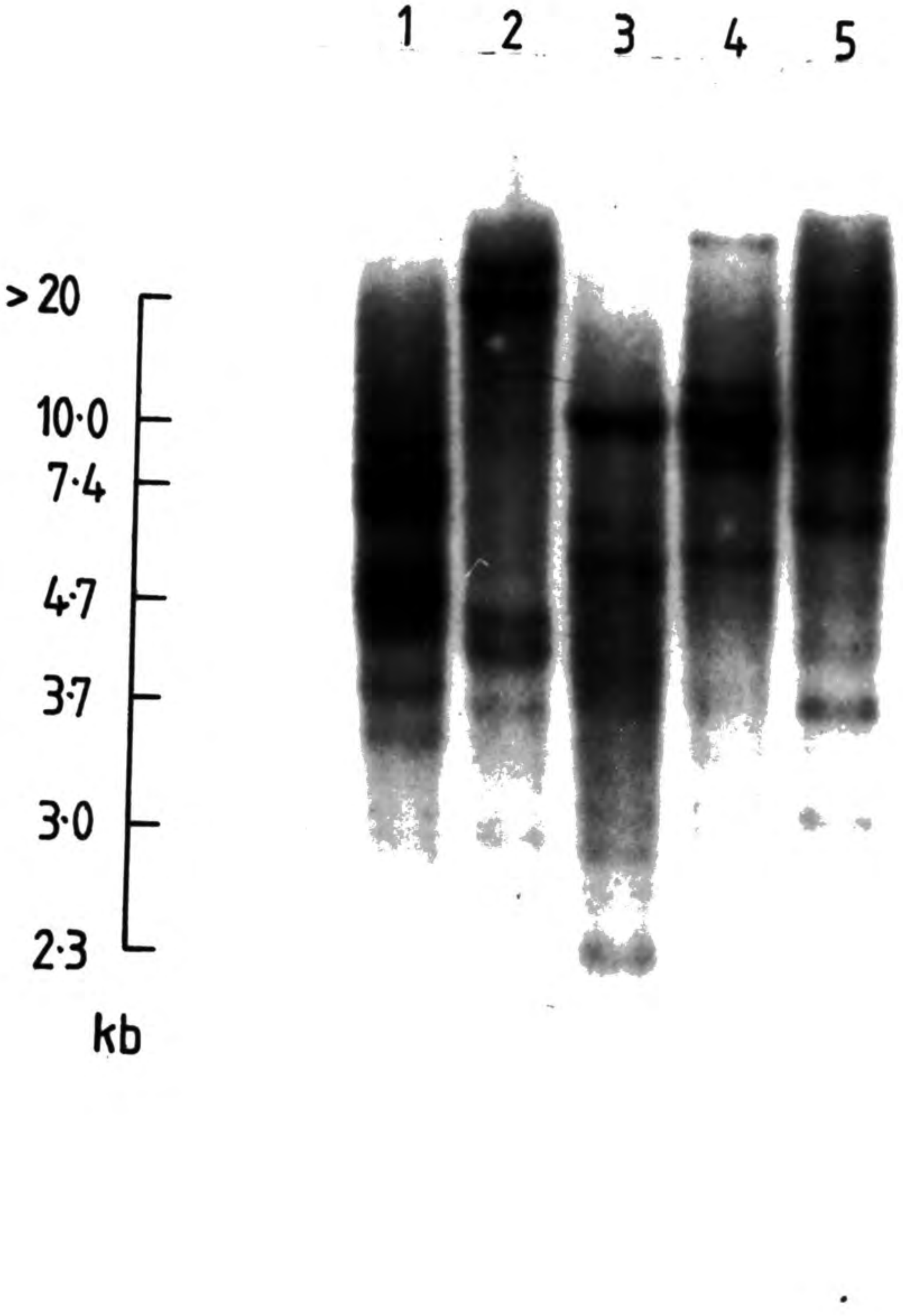

1
2
3
4
5
18·5
10·2
5·5
4·7
3·0
2·3
kb

3.3.6 Hybridisation of the 3' 0.25 Kb Eco RI/Sst I Fragment of pJC 1-16 to Genomic DNA

In order to assess whether the sequence of unknown origin at the 3' end of *Vic* B might be repeated to any degree in the pea genome, the 3' 0.25 Kb Eco RI/Sst I fragment from pJC 1-16 (see fig 14) was used as a probe specific to this region. When hybridised to a Southern blot of pea leaf genomic DNA digested with the same 5 enzymes used in section 3.3.5, the autoradiograph shown in fig. 17 resulted. Comparing this result with that achieved by hybridising only *Vic* B coding sequence to genomic DNA (see fig.16), fragments detected by the *Vic* B probe appear also with the 3' specific probe. However, this probe does appear to have hybridised to several other bands in all 5 tracks. Of possible interest also is a relatively intense 7.4 Kb band appearing in the Eco RI digestion.

Fig. 17 Hybridisation of the 3' 0.25 Kb Eco RI/Sst I fragment from pJC 1-16 (see fig. 14) to genomic DNA

Track 1 10 ug Pea leaf genomic DNA digested with Eco RI
2 " " " " " " " Bam HI
3 " " " " " " " Hind III
4 " " " " " " " Eco RV
5 " " " " " " " Bgl II


1
2
3
4
5
>20
10·0
7·4
4·7
3·7
3·0
2·3
kb

3.4 Characterisation and Partial Nucleotide Sequence of a Second Vicilin Gene, *Vic* C, also coding for a 50,000 Mr Polypeptide

3.4.1. Characterisation

λ JC-2 is another genomic clone isolated from a gene bank constructed by the ligation of fragments produced by an Eco RI partial digestion of pea leaf genomic DNA into a bacteriophage λ vector. By displaying a different pattern of hybridisation stability to the cDNA's pCD48 (see section 3.3.1) and pCD4 (encoding a 47,000 Mr Vicilin polypeptide) than observed for λ JC-I this genomic clone appears to represent a different vicilin gene to that contained in λ JC-I (Ellis et al., 1986).

λ JC-2 contains a 12.8 Kb insert, consisting of 5 Eco RI fragments. When a Southern blot of λ JC-2 digested with Eco RI was probed with pCD 48, hybridisation occurred with Eco RI with the 4.1 Kb fragment (see fig. 18, showing a restriction map of λ JC-2 and the area of cDNA homology), which as a result was subcloned into pUC9 and termed pJC 2-7.

A strategy similar to that used in section 3.3.1 for *Vic* B was used to discover the location and orientation of this potentially different vicilin gene (termed *Vic* C) in the 4.1 Kb fragment both in pJC 2-7 and also λ JC-2. This time however, in order to try and account for the different patterns of hybridisation observed between the *Vic* C and *Vic* B genomic clones, the cDNA pDUB 2, encoding a different 50,000 MR vicilin polypeptide to that of pDUB9 (used in section 3.3.1) was used. Digestion of pDUB2 with Bgl II was shown to produce 0.677 Kb 5' and 0.17 Kb 3' coding sequence specific fragments. These were then hybridised separately to identical Southern blots containing pJC 2-7 double digests with one enzyme known to linearise the plasmid, and another selected from a range known to have 6bp recognition sites. Once located and orientated this way, restriction mapping was completed using further various restriction digests to give the map shown in fig. 19.

Fig. 18 Partial restriction map of the pea genomic clone λ JC-2. The bar represents the area corresponding to *Vic* C coding sequence and introns. Restriction sites: E = Eco RI, S = Sph I.

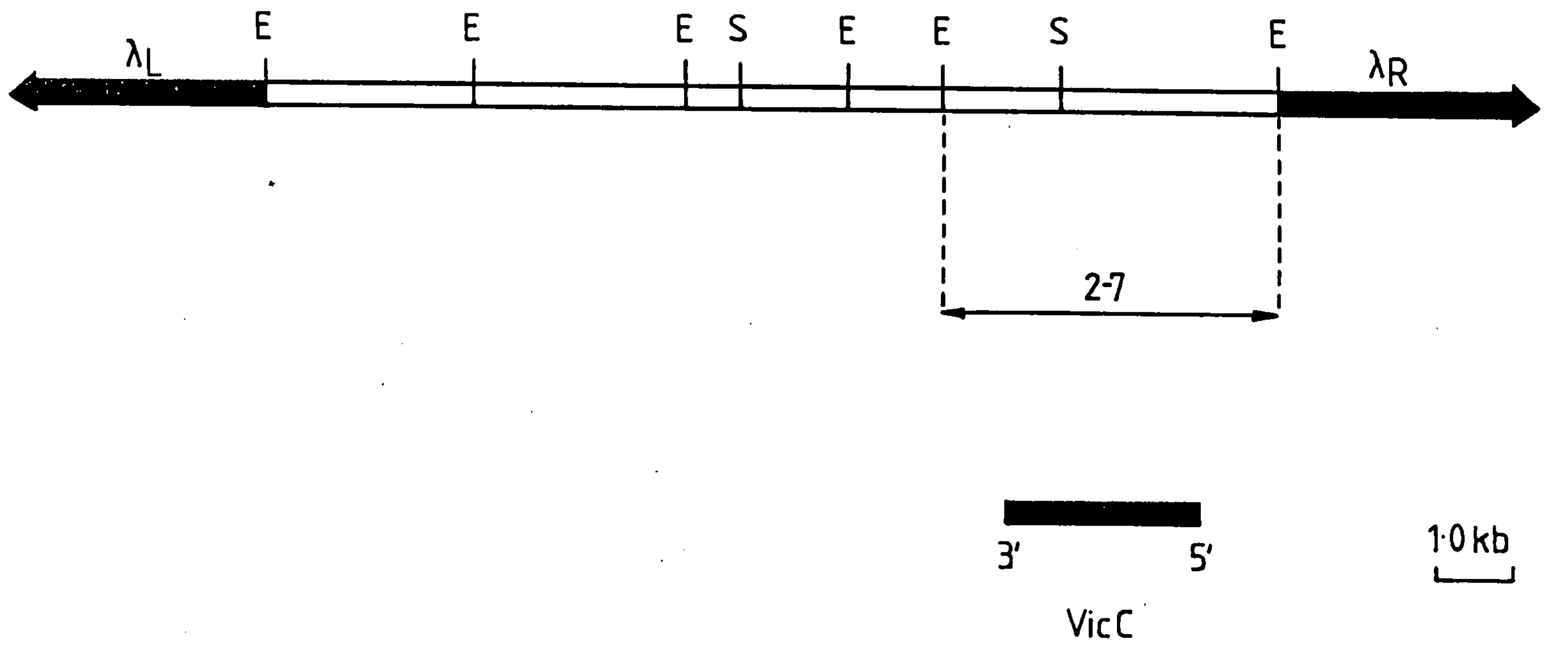
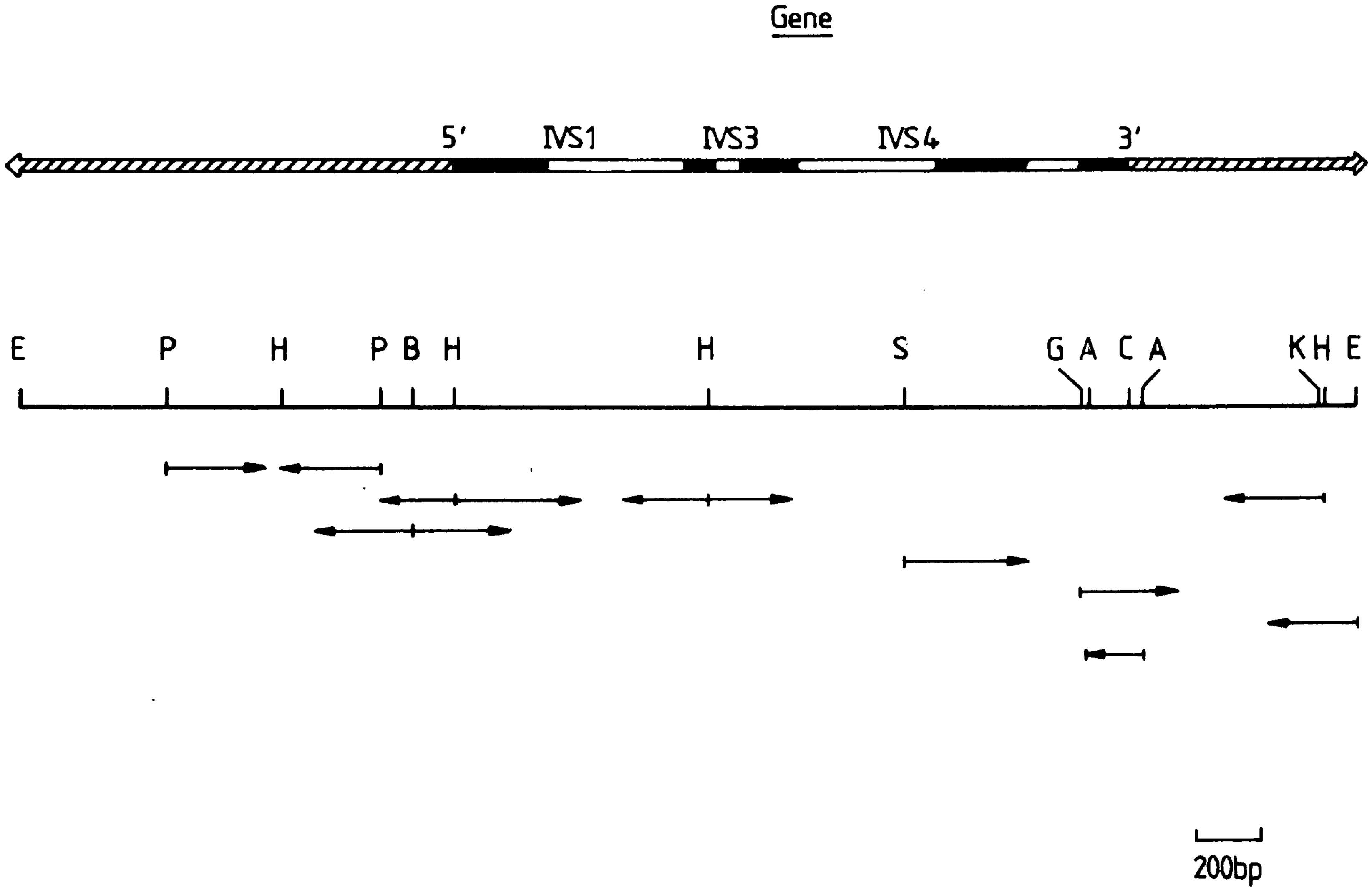

λL
E
E
E
S
E
E
S
E
λR
2·7
3'
5'
VicC
1·0 kb

Fig. 19 Restriction map and sequencing strategy for *Vic* C and its flanking regions. Restriction sites: A = Alu I, B = Bal I, C = Bcl I, E = Eco RI, G = Bgl II, H = Hind III, K = Kpn I, P = Hpa I, S = Sph I. Arrows indicate direction and length of sequence obtained from each M13 subclone. Duplicated sequencing runs carried out to check sequences are not shown. ▬ = *Vic* C coding sequence. □ = intron sequence, ▨ = flanking sequence.


Gene
5'
IVS1
IVS3
IVS4
3'
E
P
H
P
B
H
H
S
G
A
C
A
K
H
E
200bp

3.4.2 Partial Nucleotide Sequencing of *Vic* C

Nucleotide sequence corresponding to 553 bp of 5' flanking sequence (spanning the initiation codon), 151 bp of 3' flanking sequence (spanning the termination codon) and 1287 bp of coding and intron sequence is presented in fig. 20. The bulk of this data was obtained by shotgun cloning of Sau 3A, Alu I and Rsa I fragments of the 4.1 Kb Eco RI insert from pJC 2-7 into M13 sequencing vectors. Additional sequence was obtained by digesting the insert from pJC 2-7 with Hind III and Sph 1 and subcloning the fragments into an appropriately cut M13 vector. A summary of the sequencing strategey is included in fig. 19.

Fig. 20 Partial nucleotide sequence of *Vic* C and its flanking regions. The amino acid sequence is given below each line. Major features are as indicated.

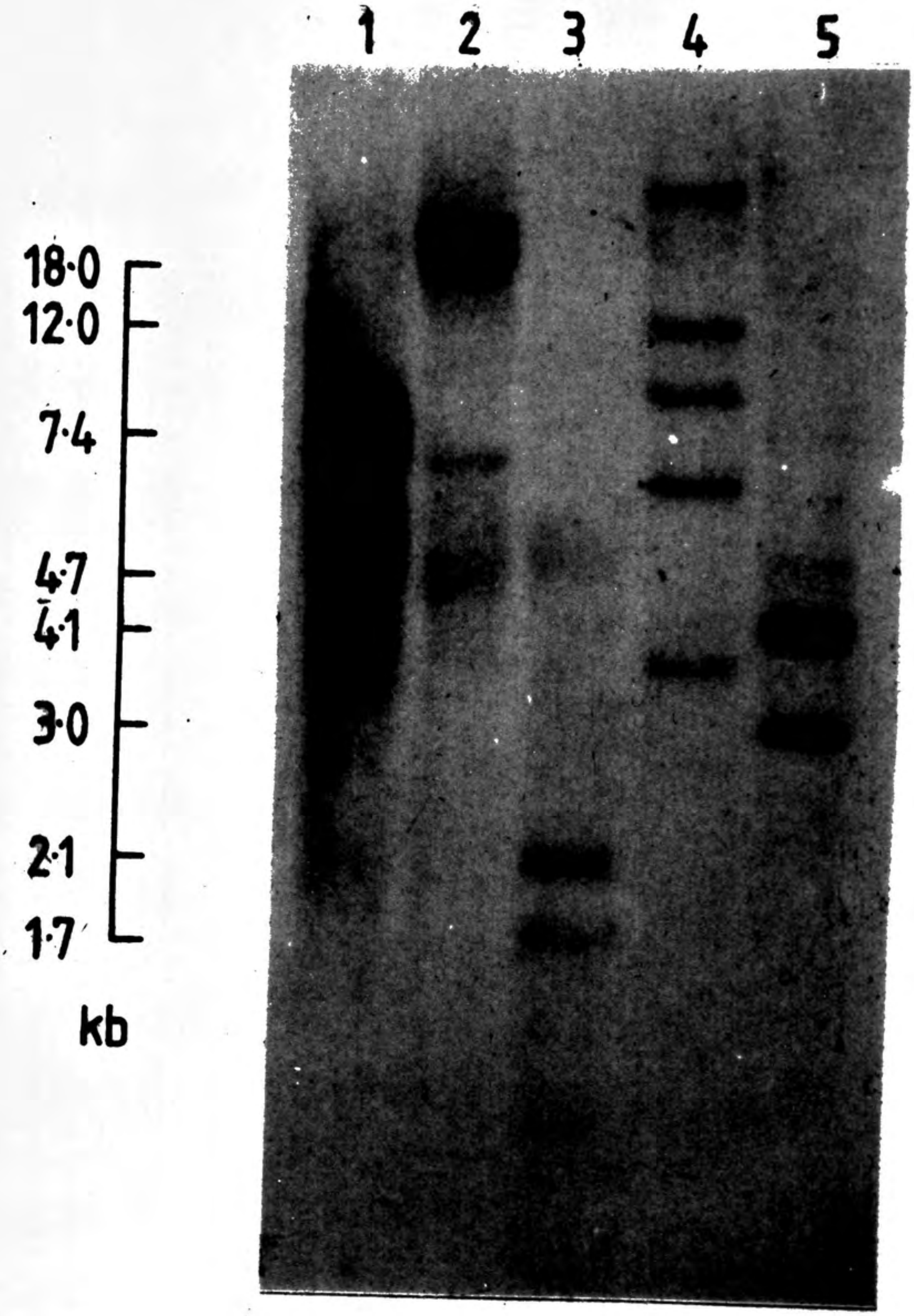
```
          ########################################
          * NUCLEOTIDE SEQUENCE OF VICLIN GENE C *
          ########################################

VicC AAAATCAAAAAGCTTTAATTAAATTTCACTAGTTGTAAATATATTTATTTATTTTTGTGTATATAATATTTATCACTTAAATTAATTTCTTAATGTTACTTTAAATTAAATTAATTCAA

VicC CCTAGTTTCGGTAGTAATGCATTTAGTAGTGAATTTTTTCAAGTCTTTATTCTCATAATAGAAGAAGTTCAGGTGAGAGACAATTTGTGGGACCCTTCATTTATTAACGTTCTCTAGCA

VicC CTTCAGAGTAATTCATAACACCAAAATCTAGCCATTGTTAACATTCTTCTTTTCCTGAAGAATGTCCAGATTCATCAAATTCAAACTTTTCTTCACCACCCATGTTATGTTCTTTTCGC

VicC TTTGCCACCTCAATTTTGTACATTTCAACACACGTCCATATGCATGGCACAACATGGCCAAATGTTGGTCATGTTAATTTATATAGCTTTCTGTTTTATACCTATAAATATCATTTGAT
     ....................................................................................................<TATA BOX>........

VicC CTCGGTGTATTTTATTCATCCAAAGTGAGTAAAGCGAGACATTAAATCAAATTAACATGGCTGCTGCTACAACAATGAAAGCTTCATTTCCACTTTTGATGGTGCTGGGAATTGCTTTC
A.A. ...............................................................................M  K  A  S  F  P  L  L  M  V  L  G  I  A  F

VicC CTAGCTTCAGTATGTGTTTCTTCTAGATCCGATCAAGATAACCCATTTATCTTTGAGTCTAAAAGGTTCCAAACTCTTTTCGAGAATGAAAAGGTTCACATTCGTCTTCTCCAAAAGTT
A.A   L  A  S  V  C  V  S  S  R  S  D  Q  D  N  P  F  I  F  E  S  K  R  F  Q  T  L  F  E  N  E  K  V  H  I  R  L  L  Q  K  F

VicC TGATCAGCGTTCTAAAATATTTGAGAATCTTCAAAATTACCGTCTTTTGGAATATAAGTCAAAACCTCACACCATATTTCTTCCACAACAAACGGATGCAGATTTCATTCTTGTTGTCC
A.A.   D  Q  R  S  K  I  F  E  N  L  Q  N  Y  R  L  L  E  Y  K  S  K  P  H  T  I  F  L  P  Q  Q  T  D  A  D  F  I  L  V  V

VicC TTAGCGGTAATTTATTATTTATCAAGTTATATATTATAAATTAATCCCTAAATCGATTATTTAAAAAAAGTATA
A.A. L  S  <-----------------------------IVS-1-------------------------------


VicC CGTTCCTCGTCAATAGCCGGTCTATTCGGTATTATAACAAGTCTTCCTTTCTTAGTATGCAACTTCATAAAATAATTTTAGAGGTCTTTTCCATAACATTATAACAACAATTGTGATCG
A.A  ............?.............<-------------------------------------------------------------------------IVS-2----------

VicC TTTTTAATCATAATTTACGACAAAAATTTCTATTGATATTATAATTATTTTTTGTCATTTCAACAGTCTTTCTTATTGTCTGGAAATCAGAACCAACAATCTATCTTATCTGGATTCAG
A.A. ----------------------------------------------------------------------> S  F  L  L  S  G  N  Q  N  Q  Q  S  I  L  S  G  F  S

VicC CAAGAACATTCTAGAAGCTTCCTTCAATGTAAGTATAAAGCACATTCTTTCTTCATTTTTTATGTATTAGTTTGATTTACTCTGTATAAATATTCCTTTTGTCATTGTAAAATATACAG
A.A.   K  N  I  L  E  A  S  F  N<---------------------------------------IVS-3------------------------------------------>

VicC ACAGATTATGAAGAGATAGAGAAGATTCTCTTAGAAGAGCATGAGAAAGAGACACATCACAGAAGAGGCCTTAGGGATAAGAGATAACAGAGCCAAGAAAAGAATGTAATAGTCAAAGT
A.A.  T  D  Y  E  E  I  E  K  I  L  L  E  E  H  E  K  E  T  H  H  R  R  G  L  R  D  K  R  *  Q  S  Q  E  K  N  V  I  V  K  V

VicC ATCAATGAAACAAATTGAAGAATTAAGTAAAAACGCAAA
A.A.   S  M  K  Q  I  E  E  L  S  K  N  A


VicC TATAGATAAAACAACTCTCCCACACACACACACACGTATATATATATATATATATATATATTTATTTATTTATTTAGTTAGTTAGTTATTCTTTCAAATTTTATCTTAGGGGTCTCTAT
A.A. ---------------------------------------------------IVS-4-------------------------------------------------> G  S  L

VicC TGTTGCCACACTATAATTCAAGGGCCATAGTGATAGTAACAGTTAATGAAGGAAAAGGGGGCTTTGAACTTGTGGGTCAAAGAAATGAGAACCAACAAGGCTTGAGAGAAGAAGATGAC
A.A. L  L  P  H  Y  N  S  R  A  I  V  I  V  T  V  N  E  G  K  G  G  F  E  L  V  G  Q  R  N  E  N  Q  Q  G  L  R  E  E  D  D

VicC GAGGAAGAGGAGC
A.A.  E  E  E  E


VicC AGATCTTACATTCCCTGGATCAGCTCAAGAGGTTGACAGGCTACTAGAGAATCAAAAACAATCTTATTTTGCAAATGCTCAACCTCAACAAAGAGAGACAAGAAGCCAAGAAATAAAGG
A.A.   D  L  T  F  P  G  S  A  Q  E  V  D  R  L  L  E  N  Q  K  Q  S  Y  F  A  N  A  Q  P  Q  Q  R  E  T  R  S  Q  E  I  K

VicC AACATCTGTATTCAATTTTGGGGGCCTTTTAATGAGTGATCAAATATTTTGCATGTATGCTATAAAGAACTATAGCTCATAATGAGCAAGGAATAAAACATCGTTCTCTTGTACTATAA
A.A. E  H  L  Y  S  I  L  G  A  F  *.....................................................................<PolyA+>...................

VicC TTATAACTCCACCTTTCTACTATGAATAATAATCAAAGATGTTATGTGCTCCTACTATGTTTT
```

3.4.3 Hybridisation of *Vic* C to Genomic DNA

The 4.1 Kb Eco RI fragment from pJC 2-7, containing the *Vic* C gene in its entirety, was hybridised to a Southern blot of pea genomic DNA digested with the same 5 enzymes used in the corresponding experiment for *Vic* B (see section 3.3.4). The resulting autoradiograph is shown in Fig. 21 - the rather complex band patterns obtained can best be summarised as follows; The Eco RI digest giving a strongly hybridising band at 7.4 Kb, with less intense bands at 8.9, 6.9, 5.0, 4.7 and 4.1 Kb. With Bam HI relatively strong bands appear at 25 and 18.0 Kb with feinter blurred bands at 6.9, 5.0 and 4.7 Kb. With Hind III hybridisation has occurred relatively weakly at 2.1 and 1.7 Kb, with an even feinter band at 5.0 Kb. 5 bands of medium intensity are found in the Eco RV track at > 30, 12.0, 9.0, 6.2 and 3.6 Kb, and with Bgl II two relatively strong bands occur at 4.1 and 3.0 Kb, with much feinter bands at 5.5 and 4.8 Kb.

Fig. 21 Hybridisation of *Vic* C coding sequence to pea genomic DNA.

Track 1 10 ug Pea Leaf genomic DNA digested with Eco RI
2 " " " " " " " Bam HI
3 " " " " " " " Hind III
4 " " " " " " " Eco RV
5 " " " " " " " Bgl II


1
2
3
4
5
18·0
12·0
7·4
4·7
4·1
3·0
2·1
1·7
kb

3.5. Analysis of Gene Copy Number In Several Different Lines of Pea

Genomic DNA was extracted from the leaves of plants of the following pea lines; Mangetout, J.I. 81, 851, 807, 808, 809^2 ,1263 and 1552. Taxonomic status and genotype for these lines is summarised in Appendix 1.

A series of Southern blots was then prepared, in each case consisting of DNA from each pea line plus a sample from the cultivar Feltham First, all digested with the same enzyme. Alongside these tracks, an appropriate size marker was included, and also a range of gene copy equivalents corresponding to the gene used to probe the blot.(Young et al., 1981). These gene copy equivalents were calculated assuming a haploid pea genome size of 4.8 x 10^9 bp (Murray et al., 1978). Various different genes were then hybridised individually to these Southern Blots thus:

3.5.1 *Leg* A

The 2.36 Kb Hind III fragment from the cDNA clone p DUB21, which contains nearly all the *Leg* A coding sequence (Lycett et al., 1984) was labelled and used to probe Southern blots containing samples digested with either Eco RI or Hind III to give the autoradiographs shown in figs. 22 (Eco RI digests) and 23 (Hind III digests).

3.5.2 *Leg* K

Firstly, the entire 3.5 Kb Eco RI insert from the plasmid pJC 5-11, containing most of the *Leg* K coding sequence and also approx. 2.2 Kb of 3' noncoding sequence (see fig. 9), was labelled and used to probe a Southern blot containing samples digested with Eco RI. The autoradiograph obtained from this is shown in fig. 24.

Next, a probe consisting only of *Leg* K coding sequence was made by isolating the 1.3 Kb BalI/Eco RI fragment of pJC 5-11 (see fig.9). The autoradiographs resulting from probing Southern blots containing samples digested with Eco RI and Hind III are shown in figs. 25 and 26 respectively.

3.5.3 *Vic* B

The 1.95 Kb Hind III fragment from the plasmid pJC 1-16, corresponding to almost all the *Vic* B coding sequence (see fig. 14) was used to probe a Southern blot containing samples digested with Eco RI to give the autoradiograph shown in fig. 27.

Sizes and intensities of all the hybridising bands obtained on these autoradiographs are summarised in Table 5.

Table 5 Summary of hybridisation patterns for the pea line blots

ECO RI	M'Tout	JI 81	851	807	808	809^2	1263	1552	F.F.
Leg A	16.3(1)	15.6(5)	14.4(3)	17.5(2)	9.2(2)	16.3(2)	18.8(2)	15.2(1)	14.4(5)
	6.9(1)	12.0(3)	12.5(3)	12.5(3)	6.7(4)	13.3(2)	10.7(1)	13.0(2)	12.5(5)
	4.2(1)	10.5(3)	8.0(3)	10.5(1)	5.35(5)	10.5(3)	8.9(7)	10.7(5)	8.0(5)
								10.5(5)	
		8.7(2)	4.4(3)	5.1(3)	4.9.3	5.1(4)	7.35(2)	5.0(5)	4.4(5)
		4.45(4)					5.2(4)	4.2(4)	
							4.4(3)		
Leg B				12.0(1)					
	9.0(3)	14.5(1)	4.9(3)	10.75(4)	7.45(3)	10.75(3)	7.45(3)	7.45(1)	7.95(4)
		10.75(4)							
	4.0(1)	9.9(4)	3.55(4)	9.90(4)	4.49(3)	9.9(3)	3.55(4)	3.55(4)	3.55(4)
	3.8(2)	3.55(2)	3.00(2)		3.55(2)	3.0(3)	3.0(2)	3.0(2)	3.0(2)
	3.55(2)	3.3(3)	2.55(2)		3.3(1)	2.9(3)	2.55(2)	2.55(2)	2.55(2)
	2.9(5)	1.9(5)	1.9(4)	3.55(2)	2.7(2)	1.9(4)	1.9(4)	1.9(5)	1.9(5)
				3.3(2)	1.9(4)				
				1.9(4)					
Vic B	20.5(2)	15.05(2)	16.0(1)	18.5(1)	19.0(4)	17.1(1)	11.15(1)	11.7(1)	15.0(3)
	12.35(2)	12.35(2)	9.45(2)	13.0(1)	12.35(3)	12.35(2)	8.85(1)	9.9(2)	12.35(3)
	9.05(1)	10.1(3)	7.8(3)	11.15(2)	7.5(1)	10.6(3)	7.5(4)	8.85(2)	8.5(3)
	7.25(1)	8.0(3)	5.85(4)	8.85(1)	4.7(2)	9.9(2)	5.85(4)	7.5(4)	7.5(4)
	3.9(4)	5.85(4)	4.7(5)	7.8(3)	3.7(2)	8.5(2)	4.7(5)	6.5(3)	6.7(3)
		4.7(5)	4.2(2)	7.0(1)	3.1(2)	7.8(3)	4.1(3)	4.7(3)	5.85(4)
		4.2(3)	4.1(2)	5.85(4)	1.9(3)	5.85(4)	3.7(1)	4.2(3)	4.7(5)
				4.7(5)		4.7(5)			4.2(3)
				4.2(3)		4.2(3)			3.7(1)
									3.5(1)

Leg K + 3' noncoding:
multiple bands obtained in each track, too numerous to give details of each - see Discussion.

Hind III									
Leg A -	11.0(3)	7.0(2)	8.0(2)	7.0(1)	-15?(1)	9.1(2)	4.7(4)	7.7(2)	6.0(1)
	7.7(2)	5.7(1)	4.7(4)	4.65(4)	6.9(4)	5.7(1)	2.8(5)	5.7(2)	4.7(5)
	6.9(1)	4.65(4)	2.8(5)	2.8(5)	2.8(5)	4.65(4)	1.85(3)	4.7(4)	2.8(5)
	5.1(4)	2.8(5)	1.6(3)	1.85(3)	1.85(3)	2.8(5)		2.8(5)	1.6(3)
	2.8(5)	1.85(4)				1.85(3)		1.85(4)	
	1.85(4)								
Leg K	11.5(3)	13(5)	11.5(4)	13(4)	11(4)	13(4)	11(4)	11(4)	11(5)
	10.3(2)	8.3(4)	10.3(4)	8.3(4)	10.3(4)	12.5(1)	10.3(4)	10.3(4)	10.3(4)
	10.2(2)	5.6(4)	8.3(2)	5.6(4)	8.3(3)	8.3(4)	8.3(2)	8.3(2)	8.3(3)
	8.3(3)	4.7(3)	6.6(3)	4.7(2)	2.5	6.6(2)	6.6(4)	5.6(2)	6.6(5)
	5.45(2)	2.5	5.6(2)	2.5		5.6(4)	5.6(2)	4.3(4)	5.6(3)
	5.3(3)		2.5			2.5	2.5	2.5	2.5
	2.5								

Hybridising band sizes are in kilobases
Figures in brackets refer to intensity of hybridisation, on a scale from 1 to 5, with 1 strongest and 5 the weakest.

Fig. 22 Hybridisation of *Leg* A coding sequence to genomic DNA from various pea lines digested with Eco RI.

Track	1	10 ug	Mangetout	leaf	genomic	DNA
	2	"	J.I. 81	"	"	"
	3	"	851	"	"	"
	4	"	807	"	"	"
	5	"	808	"	"	"
	6	"	809^2	"	"	"
	7	"	1263	"	"	"
	8	"	1552	"	"	"
	9	"	Feltham First	"	"	"

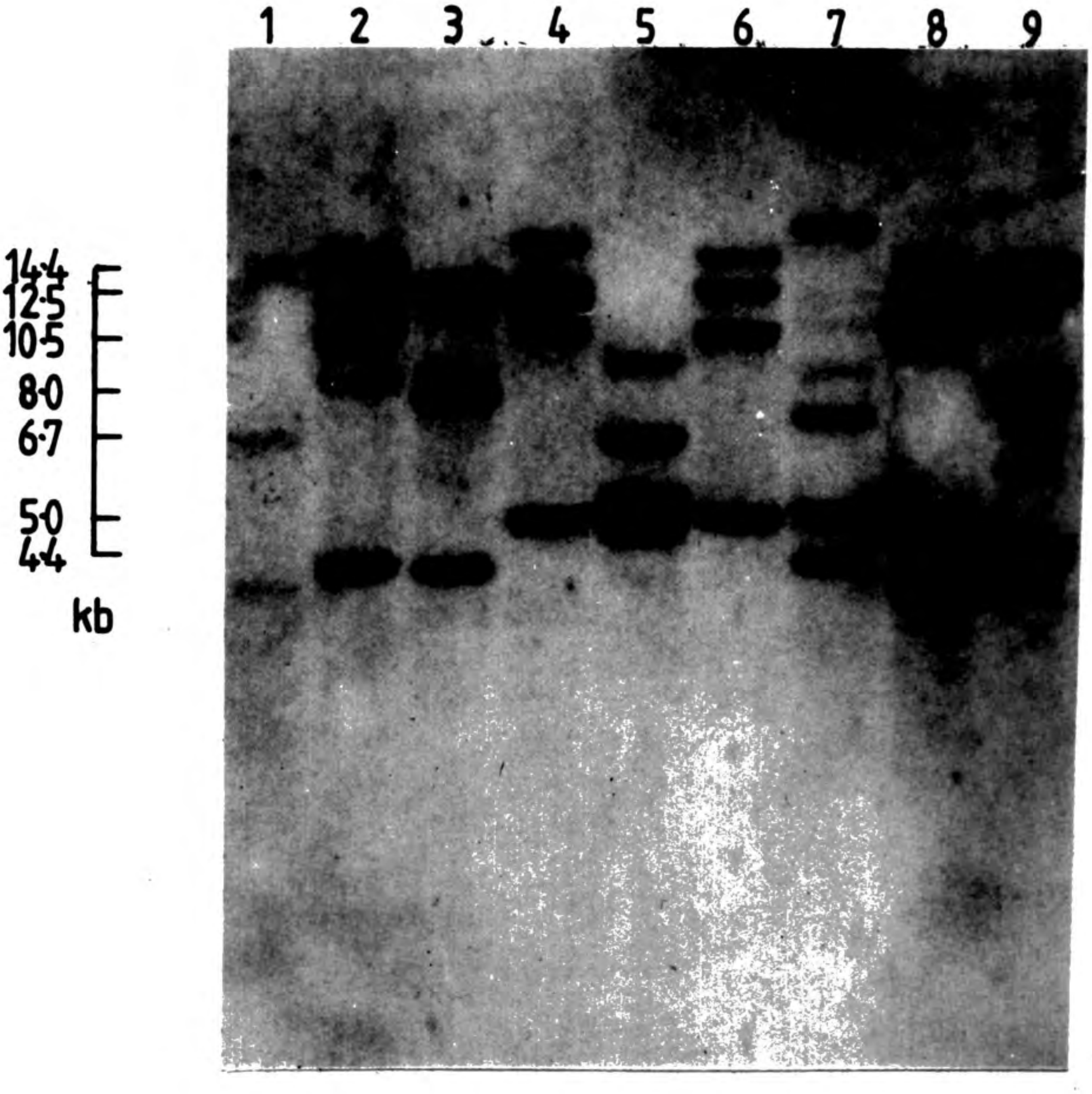

1
2
3
4
5
6
7
8
9
14·4
12·5
10·5
8·0
6·7
5·0
4·4
kb

Fig. 23 Hybridisation of *Leg* A coding sequence to genomic DNA from various pea lines digested with Hind III.

Track 1	10 ug	Mangetout	leaf	genomic	DNA
2	"	J.I. 81	"	"	"
3	"	851	"	"	"
4	"	807	"	"	"
5	"	808	"	"	"
6	"	809^{2}	"	"	"
7	"	1263	"	"	"
8	"	1552	"	"	"
9	"	Feltham First	"	"	"

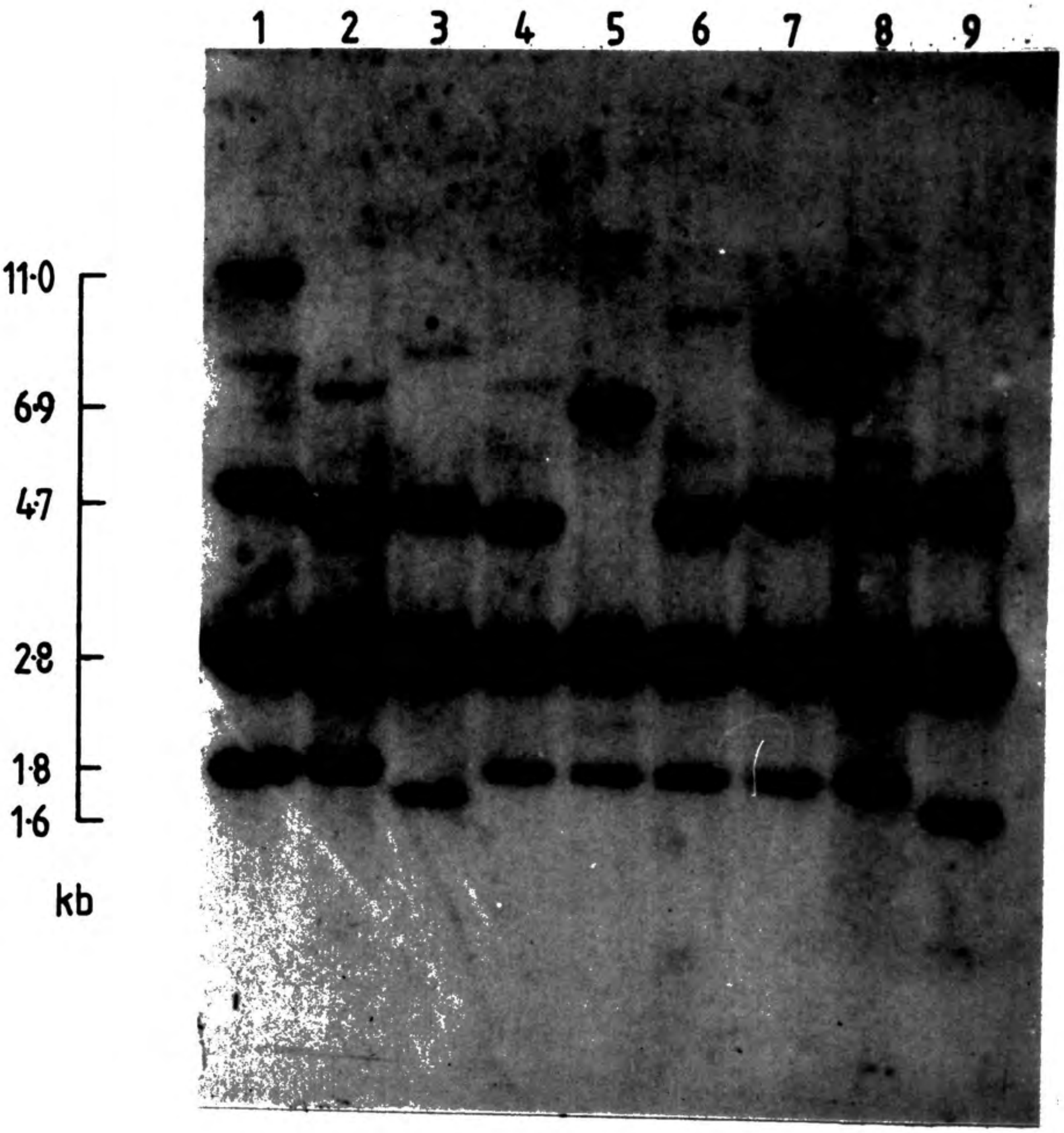

1
2
3
4
5
6
7
8
9
11·0
6·9
4·7
2·8
1·8
1·6
kb

Fig. 24 Hybridisation of *Leg* K 3' flanking sequence to genomic DNA from various pea lines digested with Eco RI.

Track 1	10 ug	Mangetout	leaf	genomic	DNA
2	"	J.I. 81	"	"	"
3	"	851	"	"	"
4	"	807	"	"	"
5	"	808	"	"	"
6	"	809^2	"	"	"
7	"	1263	"	"	"
8	"	1552	"	"	"
9	"	Feltham First	"	"	"

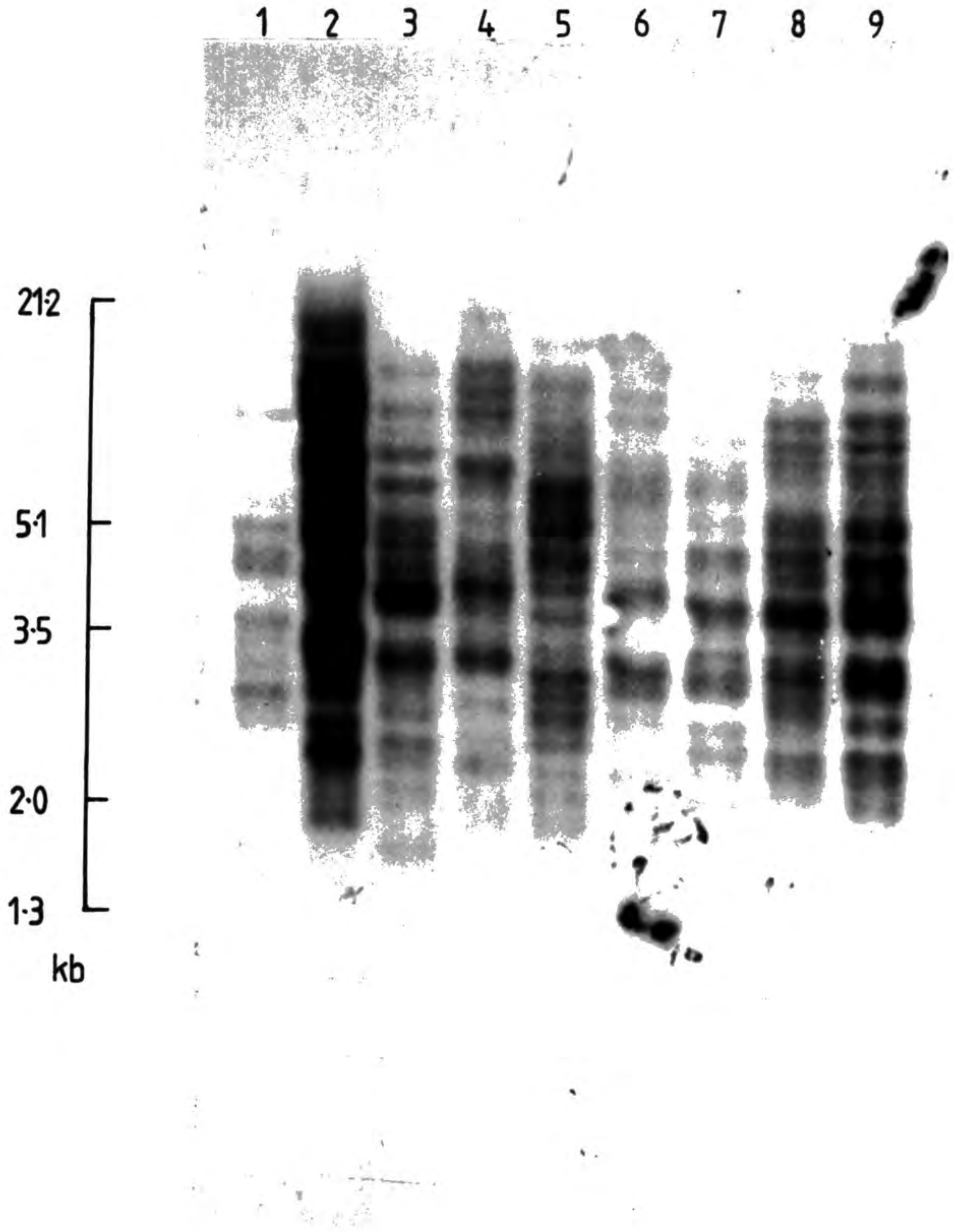

1
2
3
4
5
6
7
8
9
21·2
5·1
3·5
2·0
1·3
kb

Fig. 25 Hybridisation of *Leg* K coding sequence to genomic DNA from various pea lines digested with Eco RI.

Track	1	10 ug	Mangetout	leaf	genomic	DNA
	2	"	J.I. 81	"	"	"
	3	"	851	"	"	"
	4	"	807	"	"	"
	5	"	808	"	"	"
	6	"	809^2	"	"	"
	7	"	1263	"	"	"
	8	"	1552	"	"	"
	9	"	Feltham First	"	"	"
	10	1	*Leg* K	gene	copy	equivalent
	11	2	"	"	"	equivalents
	12	5	"	"	"	"

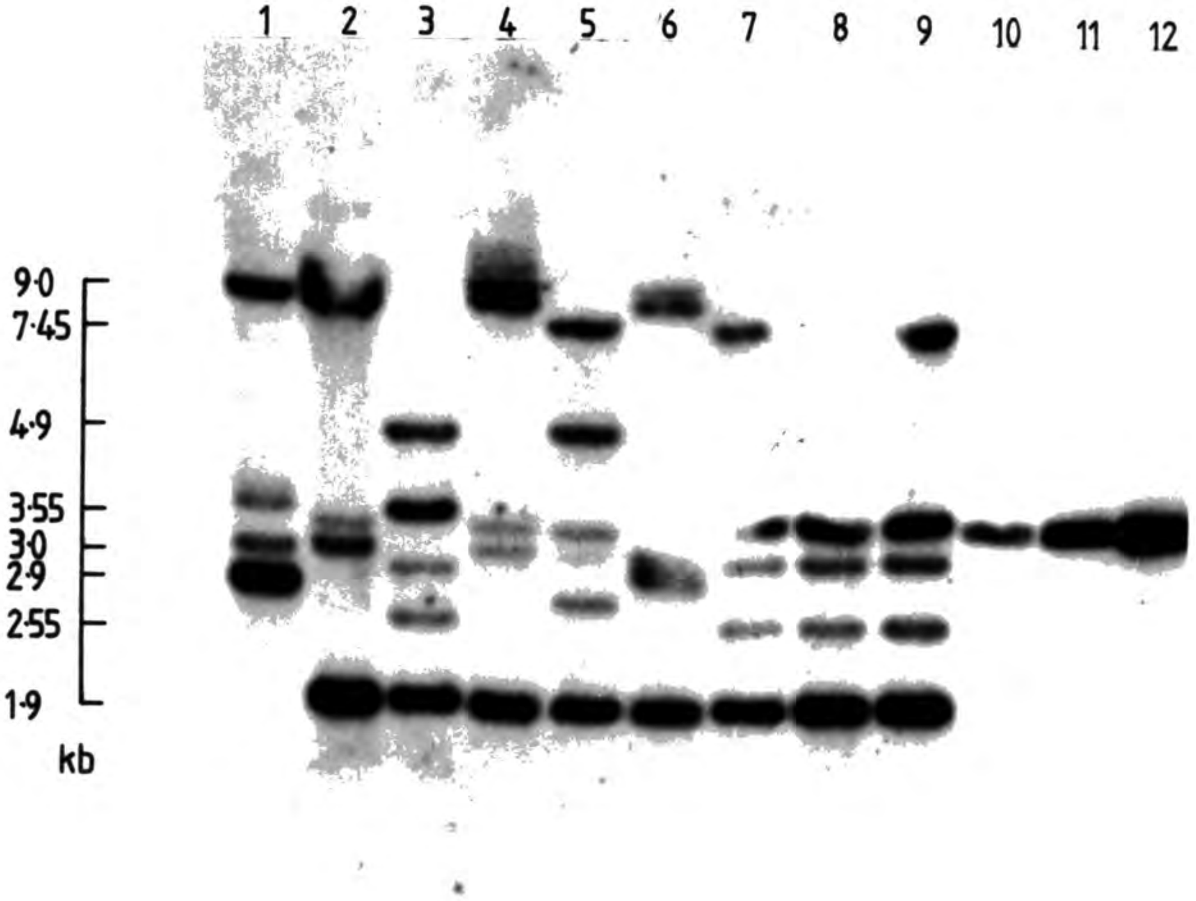

1
2
3
4
5
6
7
8
9
10
11
12
9·0
7·45
4·9
3·55
3·0
2·9
2·55
1·9
kb

Fig. 26 Hybridisation of *Leg* K coding sequence to genomic DNA from various pea lines digested with Hind III.

Track	1	10 ug	Mangetout	leaf	genomic	DNA
	2	"	J.I. 81	"	"	"
	3	"	851	"	"	"
	4	"	807	"	"	"
	5	"	808	"	"	"
	6	"	809^2	"	"	"
	7	"	1263	"	"	"
	8	"	1552	"	"	"
	9	"	Feltham First	"	"	"

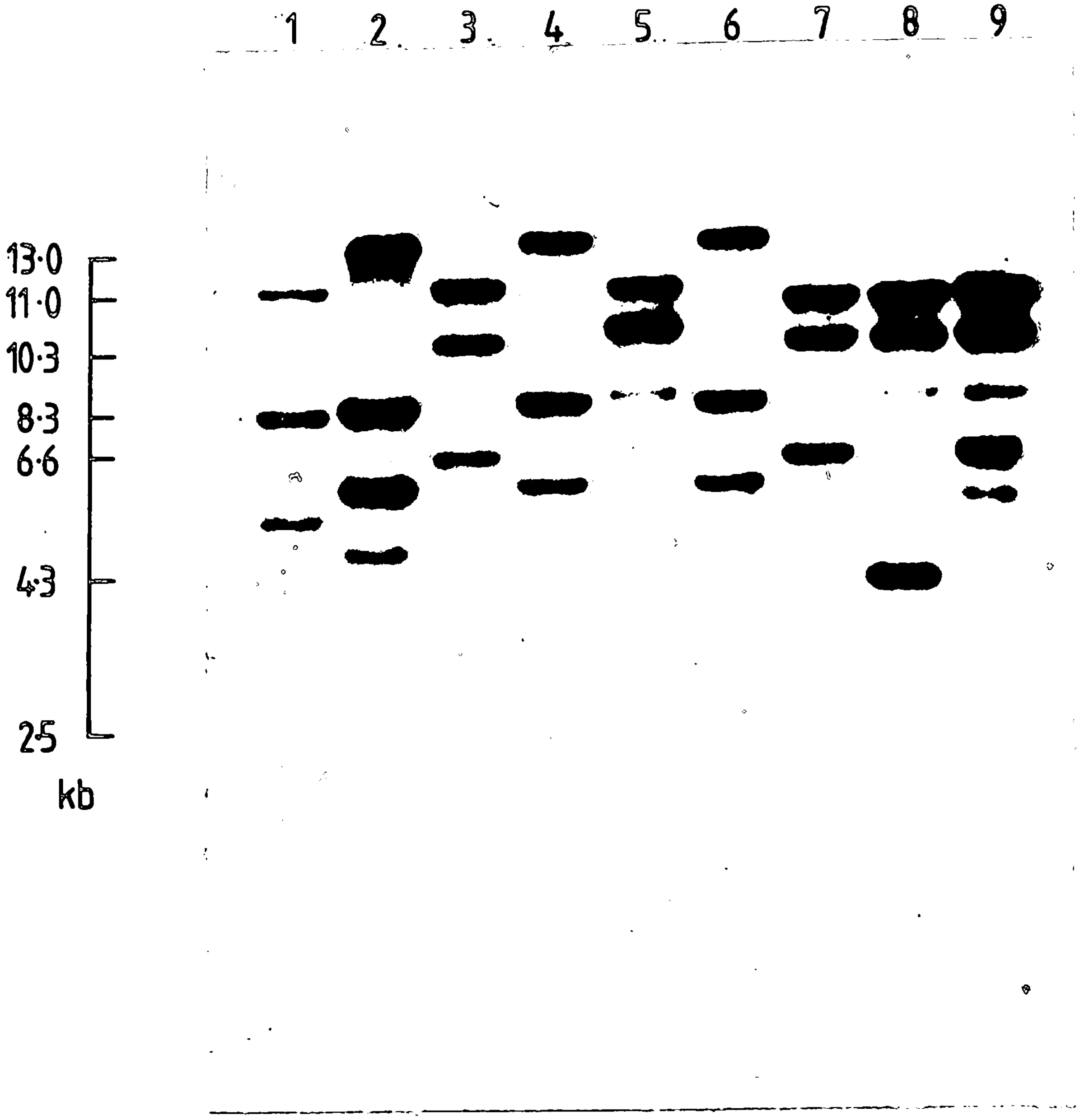
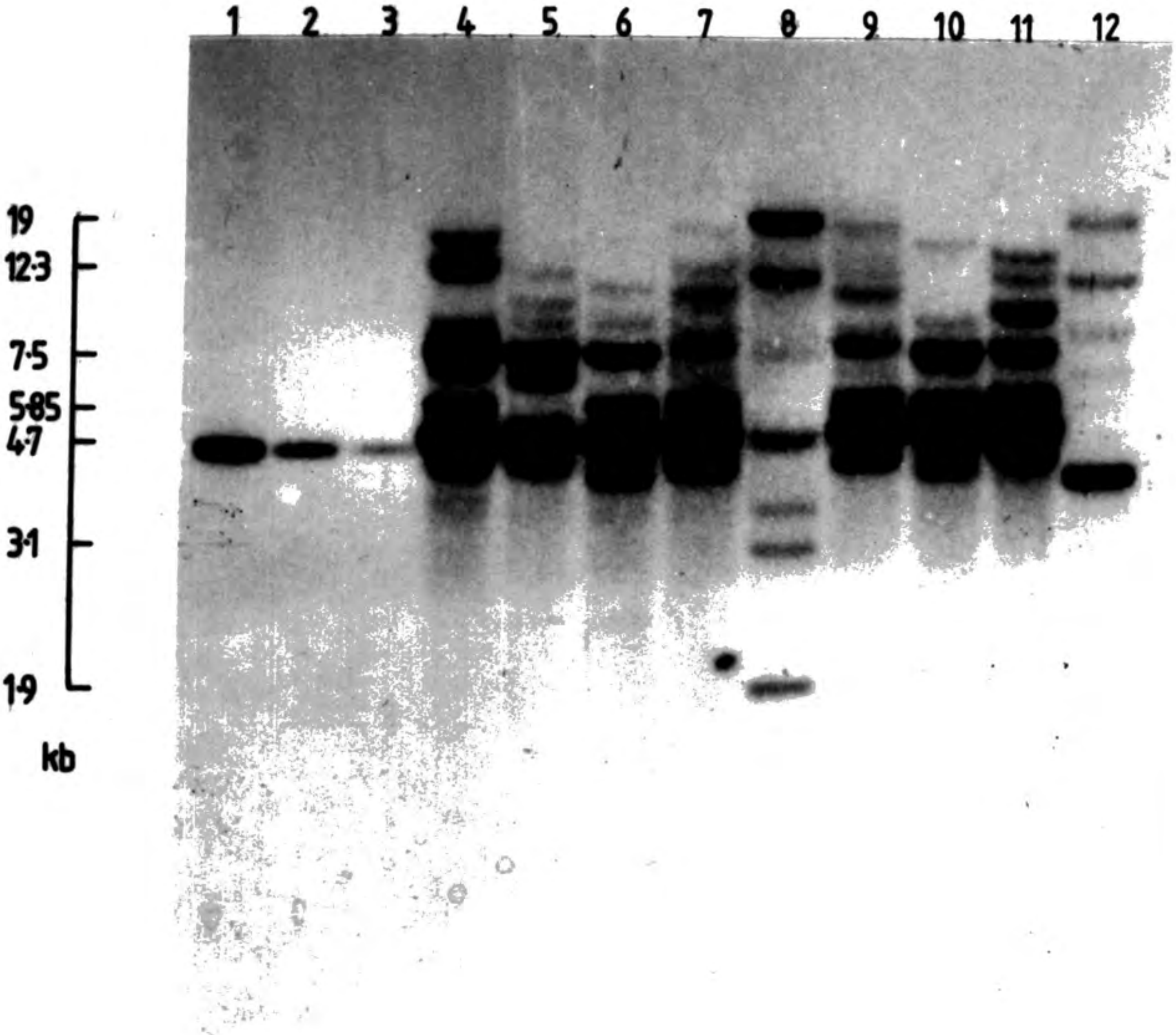

1
2
3
4
5
6
7
8
9
13·0
11·0
10·3
8·3
6·6
4·3
2·5
kb

Fig. 27 Hybridisation of *Vic* B coding sequence to genomic DNA from various pea lines digested with Eco RI.

Track 1 5 *Vic* B gene copy equivalents
2 2 " " " "
3 1 " " " equivalent.
4 10 ug Feltham First leaf genomic DNA
5 " 1552 " " "
6 " 1263 " " "
7 " 809^2 " " "
8 " 808 " " "
9 " 807 " " "
10 " 851 " " "
11 " J.I. 81 " " "
12 " Mangetout " " "


1
2
3
4
5
6
7
8
9
10
11
12
19
12·3
7·5
5·85
4·7
3·1
1·9
kb

CHAPTER 4

DISCUSSION

4.1 Storage Protein Gene Structure

4.1.1. *Leg* D

The most significant feature of *Leg* D emerges from the comparison of its nucleotide sequence with that of *Leg* A (fig.3), which reveals, due to certain crucial differences, that *Leg* D appears to be a pseudogene, (see section 1.1.4.)

Several observations confirm this; firstly the presence of two in-frame stop codons in the coding sequence of *Leg* D, corresponding to amino acid positions 6 and 29 in *Leg* A. Secondly, there is a single base pair deletion in *Leg* D, (at amino acid 150 in *Leg* A) which would lead to an error in the reading frame were the gene to be transcribed and translated. Finally, *Leg* D contains 3 sizeable deletions; i) the 3' intron/exon boundary of IVS-2 along with amino acids 159-205 of *Leg* A (coupled with additional sequence in IVS-2 in *Leg* D), ii) the region corresponding to amino acids 251-272 in *Leg* A, iii) amino acids 406-439 of *Leg* A resulting, due to an extra base in *Leg* D, to another frameshift. From this evidence, and also the observation that *Leg* D retains the 3 introns found in *Leg* A in corresponding positions, it is clear that *Leg* D is a non-processed type pseudogene (see section 1.1.4)

As far as basic homology between the two genes, there is close agreement at the nucleotide sequence level, with 150 base substitutions over comparable regions of coding sequence (equivalent to 90.3% homology over the 1554 bases of coding sequence). Of these base substitutions, 69 result in amino acid changes ('replacement' substitutions), with the remaining 81 giving no alteration in amino acid sequence ('silent' substitutions). Further evidence of the nonfunctional nature of *Leg* D might be implied by a comparison of its leader sequence with that of *Leg* A, B and C (Lycett et al., 1985) all of which appear to be members of a closely conserved gene family. This shows no changes in amino acid sequence in the 21 amino acid residue leader between the 3 functional genes, *Leg* A, B and C, but 4 amino acid changes, the insertion of a Threonine residue (betwen a.a.s -20 and -19 in *Leg* A), and deletion of a Leucine residue (-17 in *Leg* A) in the leader region of *Leg* D. For a region which would normally be expected to be under considerable functional constraint, the differences shown in *Leg* D suggest that this functional constraint no longer exists since *Leg* D has ceased to code for a mature polypeptide, and mutations in this region are no longer selected against.

Examination of the 5' flanking sequence of *Leg* D also reveals some interesting features. A 'TATA' box is present in a comparable position to *Leg* A, but homology between the CAAT box of *Leg* A and the comparable region in *Leg* D appears to have broken down. Homology around the transcription start site has been conserved, but in *Leg* D the site is positioned differently relative to the start codon due to a 7 base pair deletion downstream of transcription start. As far as the putative regulatory element specific to legumin genes, the 'legumin box' (Baumlein et al., 1986) is concerned, present in *Leg* A, B and C as a 29 bp region of homology starting at -118 (relative to the transcription start site), the corresponding region in *Leg* D shows 3 base changes, only 2 of which differ from the 'box' identified by

Baumlein using *Vicia faba* Legumin gene *LE B* $_4$, *Leg* A from *Pisum Sativum* and G1, a glycin in gene from *Glycine max*, and thus apparently not yet differing significantly from this conserved sequence. Generally, using dot matrix comparison (results not presented), sequence homology between the 5' flanking regions of *Leg* D and *Leg* A breaks down upstream of position -145 in *Leg* D. Consequently, the region of 8bp showing 90% homology to the SV40 enhancer core sequence (reported by Weiher et. al., 1983) observed in *Leg* A, B and C (Lycett et al., 1985) is also present in Leg D. However, further upstream, and now in the region where homology between Leg D and Leg A has broken down, the 10bp region showing 80% similarity to the adenovirus enhancer core element (Hearing and Shenk 1983) present in *Leg* A, B and C (Lycett et al., 1985) is not observed in *Leg* D.

The 3' flanking sequence of the two genes shows complete divergence after the stop codon. *Leg* D has 3 potential polyadenylation signals, but in completely different positions to the 3 in *Leg* A. Also, in Leg D the 1st and 2nd signals are overlapped, whereas in Leg A this occurs with the 2nd and 3rd signals.

Sequence divergence between the introns of *Leg* A and D varies. Intron 1 shows greatest conservation with 49/88 bases remaining identical (55.7% homologous). Only part of the 2nd intron is comparable, with 34/40 bases at the 5' end being the same, then followed by complete divergence. Intron 3 shows even less similarity, with only 3 bases at the 5' end and 15 at the 3' end being the same (see section 4.1.6 for potential significance). These observations confirm the view that intron sequence will diverge more rapidly than coding sequence between two comparable genes due to lower functional constraint (Shah et al., 1982). Additionally it can be observed that the intron/exon boundaries of *Leg* D (where present) obey the Breathnach-Chambon rule (Breathnach et al., 1978).

4.1.1.1 Hybridisation of *Leg* D to Genomic DNA

Hybridisation of the coding sequence probe for *Leg* D to restriction enzyme digests of genomic DNA detected various fragments, as shown in fig 6. The probe detected its matching fragment in the Hind III digested DNA, at 1.5 kb, but at much lower intensity than was obtained for the single copy *Leg* D standard. The probe also hybridised strongly to the 12.5 kb Eco R1 fragment (containing both *Leg* D and *Leg* A, represented in the λ Leg 1 genomic clone - see section 3.1.1), and to the 2.4 kb Hind III and 3.5 kb Bam H1 fragments corresponding to the equivalent fragments from λ Leg 1, containing *Leg* A (see fig 1). A similar result was obtained with the 5' flanking sequence probe (see fig 7), with the matching 0.8 kb Hind III fragment being detected at lower intensity than was the single copy equivalent. Such disparity in hybridisation intensities observed between that detected for the gene fragment in the genomic DNA and that for a single gene copy equivalent might possibly be put down to poor transfer of smaller fragments relative to larger ones in the genomic DNA during Southern blotting, (as suggested in the stronger intensities of other, larger, bands containing *Leg* D or *Leg* A, which are apparently present at approx single copy level in the genome). Alternatively, an explanation might be incomplete digestion of the genomic DNA into all its constituent fragments - however as no larger bands of strong intensity were present other than those already accounted for, plus the fact that the genomic DNA digests appeared complete on the gel (see fig. 5), this hypothesis is unlikely.

4.1.2 Leg K

Alignment and comparison of the nucleotide sequences of *Leg* K and *Leg* J (fig. 10) shows that unfortunately the 5' end of *Leg* K, including both flanking sequence, the leader sequence, and 154 amino acid residues from its predicted N terminal end are truncated by the genomic clone λ JC5. The predicted amino acid sequence from *Leg* J corresponds to N-terminal sequences identified in 'minor' legumin polypeptide species (Gatehouse et al., 1988). The two genes show a high level of homology between their coding regions, with 1023 out of 1050 bases matching (=97%). The differences in coding sequence between *Leg* J and *Leg* K give a total of 5 in-frame codon-deletions (4 in *Leg* K, 1 in *Leg* J, with the remaining 27 base changes due to 16 'silent' substitutions and 11 'active' substitutions.

Comparison of sequences between the gene and homologous cDNAs shows the two introns to be in corresponding positions. In *Leg* J, they are 138 and 98 bases long, in *Leg* K, 81 and 105. Highest conservation of sequence is shown at the 3' ends of the introns (see also section 4.1.6) Homology at the 5' ends is lower, especially in intron 2 Intron/Exon boundaries are faithful to the rule of Breathnach et al., 1978. Overall, homology across the introns is 56% for Intron 1 and 71% for Intron 2, however if deletions are excluded, these figures rise to 96 and 74% respectively. Such a high degree of homology between introns is very significant, and points very strongly to the genes having only recently diverged.

Further evidence of recent divergence of the two genes comes from a comparison of 3' flanking sequence, where in the first 217 bases after the stop codon, there are 163 matching bases (75% homology), ignoring deletions. Homology seems to break down in the final 60 bases of *Leg* K, with 45/60 mismatches. Both genes have at least 4 potential polydenylation signals (possibly 5 in *Leg* K) in corresponding positions, with the first overlapping in *Leg* K. Using data from a cDNA homologous to *Leg* J (Domoney et al., 1987), the indication is that the 2nd or 3rd signals are likely to be used, since this cDNA shows a poly A tail in a position equivalent to 1935 in Leg J.

4.1.2.1 Sequence Comparisons

To achieve a more comprehensive picture, sequence comparisons with other storage protein genes were made using *Leg* J, rather than *Leg* K, owing to lack of sequence data from the 5' end of *Leg* K.

i) *Leg* J vs *Leg* A

Alignment of these two sequences shows a 48% level of homology, consequently giving a significant degree of similarity between the resulting amino acid sequences: This homology appears to be greatest in the N-terminal half of the α -subunit (excluding the 8 N-terminal residues).

Differences in nucleotide sequence result in 66 codon deletions between the two genes, plus 2 single base pair deletions causing alteration of the reading frame over short regions.

Two large deletions occur in the C-terminal end of the α subunit

- one of 19 codons in *Leg* A with respect to *Leg* J, and one of 35 codons in *Leg* J corresponding to the region of sequence repeats in *Leg* A. Indeed, this repeated region displays extensive divergence despite a broad overall similarity in nucleotide and amino acid composition. Any matching of codons in this region show replacement substitutions outnumbering silent ones although often the resulting amino acid changes are conservative (ie changes result in amino acids of similar charge properties).

When comparing the low level of homology shown between *Leg* J and *Leg* A, against the much higher level shown by *Leg* J and *Leg* K, it is apparent that divergence of the two gene subfamilies must have happened much earlier in evolution than did the divergence of genes within the subfamilies.

A comparison of the non-coding regions of *Leg* J and *Leg* A shows no apparent homology in the 3' flanking sequences (even the polydenylation signals differ in their positioning with respect to the termination codon). Corresponding introns (namely IVS-1 of *Leg* J vs IVS-2 of *Leg* A, and IVS-2 of *Leg* J vs IVS-3 of *Leg* A) show very low levels of homology, even when deletions are ignored. Indeed any homology shown must be of very limited significance since certain regions (boundary regions, branch points) are bound to show similarity due to considerable constraint operating for functional reasons (see section 4.1.6). It is interesting to note that there is no marked divergence in sequence between the two genes around the first intron in *Leg* A (which is absent in *Leg* J), suggesting that the gain (or loss) of IVS-1, and the evolution of intron sequences in general seems to be a process which has occurred independent of coding sequence evolution.

The 5' flanking sequences show a high level of homology (50%) relative to the rest of the areas of the genes compared. Strongest conservation occurs around the 'TATA' box and a region immediately 5' to the CAAT box (see below). This, however, is not the case around the putative 'enhancer' sequence elements observed in *Leg* A by Lycett et al., 1985. The strongly conserved region (25/28 bases) 5' to the 'CAAT' box at -80 to -108 in *Leg* J, -90 to -118 in *Leg* A, corresponds to that observed by Baumlein et al., 1986 for several other storage protein genes (see also section 4.1.1). An additional sequence with as yet unknown significance is a stretch of 10 bases from -247 to -238 in Leg J, AGGGGACCAT, also present in Leg A, but in the opposite orientation, running from -235 to -226, notable as the area in which homology between the sequences disappears. (Comparisons are summarised in Table 6)

Figure 28 shows 2-dimensional representations of the two genes (Gates, 1985) with the 4 bases being plotted as unit distances of the 4 vectors +x, -x, +y, -y (see figure for details), and clearly displays not only the A-G rich central sections of the genes (coding for the hydrophilic C-terminal ends of the α subunits), but also the overall general level of homology. The two introns common to Leg J and Leg A border the A-G rich central section, which may suggest an agreement with the theory of introns defining functional domains of the protein structure. Stone et al., 1985, suggested a similar role for introns in the evolution of the chicken Glyceraldehyde Phosphate

Dehydrogenase gene, and 2 reports by Sudhof et al., 1985a, 1985b, proposed that by defining functional domains encoded by exons, the presence of introns allowed shuffling of these domains to take place, and, in the case of the Low Density Lipoprotein receptor and Epidermal Growth Factor precursor in humans, could have facilitated the 'recruitment' of important functional domains from other genes (see section 4.1.7 for further discussion).

TABLE 6 Homologies between Leg A and Leg J, and Leg K and Leg J at the nucleotide level. Substitution and deletions are given in bases; identical bases = total bases - (substitutions + deletions).

	LEGA → LEG J				LEGK → LEG J			
	TOTAL BASES•	SUBSTITUTIONS	DELETIONS	HOMOLOGY%	TOTAL BASES	SUBSTITUTIONS	DELETIONS	HOMOLOGY %
5' FLANKING SEQUENCE -292* → -1	292	83	64	50				
5' NON TRANSLATED	44	4	16	52				
EXON 1/1	306	114	21	56				
INTRON 1+	88	-	-	-				
EXON 2/1	258	94	6	61	101	1	0	99
INTRON 2/1	88/138	-	-	NS	138	3	57	56
EXON 3/2	710	208	200	41	588	16	18	94
INTRON 3/2	99/98	61	1	36	107	23	11	71
EXON 4/3**	401	195	26	46	384	28	6	91
3' FLANKING‡	134/132	-	-	NS	265 §	45	21	75

NS Not significant

• Allowing for deletions in one or other sequence to optimise homology. All deletions in coding sequence were made by codon except for single base deletions at 944 and 1503 (Leg J in the J→A comparison).

* No significant homology 5' to this

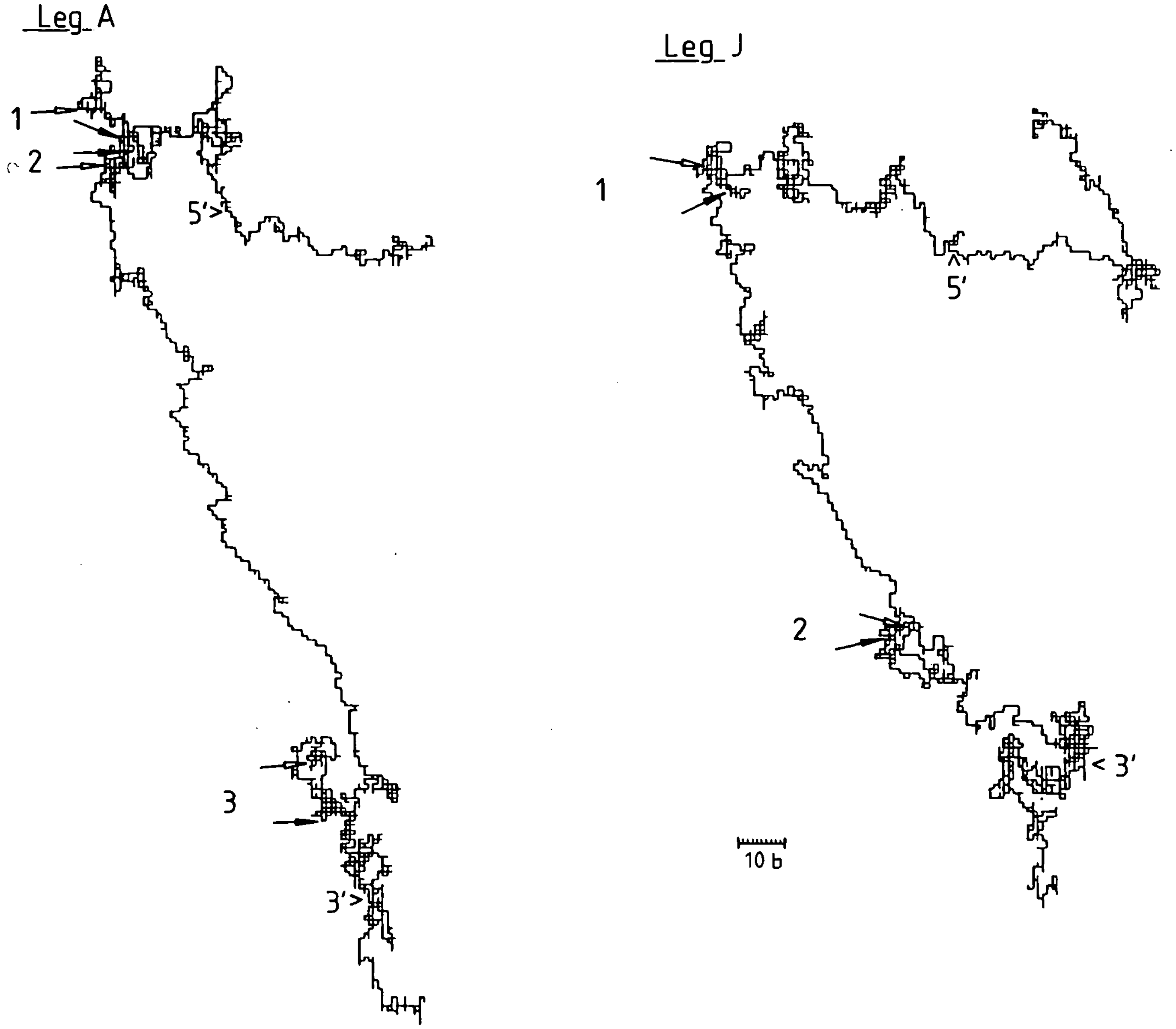
+ Absent in Leg J

‡ To known polyadenylation sites in cDNA species

§ No significant homology 3' to this

** To stop codon in Leg A

Fig 28 Comparison of coding and immediate flanking nucleotide sequences of *Leg* A and *Leg* J by the 'dimensional plot' method (Gates, 1985). A is represented as a vector (0, -1), C as (-1, 0), G as (1, 0) and T as (0, 1); the plot is produced by joining the end points of the resultant vectors after each base is added to the sequence. Introns have been omitted for clarity, but their positions are indicated by corresponding numerals. The ends of the coding sequence (5' and 3') are also indicated. Note the asymmetry and similarity of sequence composition in the highly variable a-subunit C - terminal regions of the coding sequences (between introns 2 and 3 in *Leg* A, 1 and 2 in *Leg* J).


Leg A
1
2
5'>
3
3'>
Leg J
1
5'
2
< 3'
10 b

ii) *Leg* J and *Leg* K vs *Vicia faba* Legumin Gene *LeB* $_4$

The publication of sequence for a 'B-type Legumin' gene, termed *LeB*$_4$, from *Vicia faba* (Baumlein et al., 1986) now permits detailed comparisons between this and corresponding genes from other species to be made.

Firstly it is evident that *LeB*$_4$ shares much greater homology with *Leg* J and K than it does with *Leg* A (data not presented) and one can thus conclude that *LeB*$_4$, and *Leg* J and K are members of the corresponding sub-family of genes, within their respective species.

Homology between *LeB*$_4$ and *Leg* J extends from -450 bp in the 5' flanking region downstream to 265 bp beyond the stop codon in the 3' flanking region of *Leg* J, a total of approx 2500 bp. Compared to *Leg* K, homology continues further in the 3' direction, in fact as far as the 3' sequence was determined in *Leg* K (i.e. an extra 35 bp as shown in fig. 12, plus a further 50 bp not presented) Table 7 gives data for full analyses of homology.

With respect to coding sequence, the genes show a high degree of homology, especially in exons 1 and 3 where it is in excess of 90%. Interestingly, this figure drops to approx. 75% for exon 2, which contains the C-terminal region of the α subunit and shows strong variability both in terms of deletions and substitutions. This characteristic also emerged in the comparison of *Leg* J with *Leg* A (section 4.1.2.1,p.130), clearly indicating this region to be evolving in a different way to the rest of the coding sequence of legumin genes, and as such a region under considerably less functional constraint.

Overall, the coding sequence of *LeB*$_4$ shares greater homology with *Leg* K than it does with *Leg* J, notably in better matching of deletions in the C-terminal region of the α -subunit.

With respect to non-coding sequence, the corresponding introns, of similar lengths in the *P.sativum* and *V.faba* genes show a significant degree of homology, with divergence appearing to have occurred primarily by deletion rather than substitution. The 60% level of homology shown suggests a relatively recent divergence for these *P.sativum* and *V.faba* legumin genes. Again, confirming coding sequence homology, *LeB*$_4$ is more homologous to *Leg* K than *Leg* J in IVS -2 (the fact that this is not so in IVS-1 is due to deletions rather than substitutions)

The 5' non-translated sequences of the two genes appear to be highly homologous, with both genes having the same indicated transcription start site which also appears to be conserved in *Leg* A, (Lycett et al., 1984) and G1 from Glycine max (Baumlein et al., 1986). Overall *Leg* J's 5' flanking sequence shows 80% homology to *Vfa LeB*$_4$ until -250, a higher level than that shown between introns, and suggesting some degree of functional constraint to cause this sequence conservation. Indeed, homology continues to -450 at a comparable level to that in the introns, indicating a limit for the extent of immediate 5' flanking regions of the genes.

Even more striking is the level of homology shown by the 3' flanking sequences of *Leg* K and *Vfa LeB*$_4$ - at approx 90%, it is comparable to the level shown by the coding sequences, and is significantly higher than the conservation between *Leg* J and *Leg* K in this region (section 4.1.2) (a similar situation exists in exon

3, *VfaLeB*$_4$ and *Leg* K showing slightly higher homology than *Leg* J and *Leg* K in this region) (Possible patterns of sequence divergence will be discussed in section 4.2). Since convergent evolution in the 3' flanking regions is extremely unlikely, the conservation shown in this region suggests a strong evolutionary constraint, possibly in accordance with some functional role.

iii) Sequence Comparisons of *Leg* J and *Leg* K with Soybean Glycinin cDNA Species.

The sequences of two full length glycinin cDNA species, encoding the A_3 B_4 and A_5 A_4 B_3 subunits, have been published (Fukazawa et al., 1985, and Momma et al., 1985, respectively). The cDNA species are clearly more closely related to *Leg* J, K and *Vfa LeB*$_4$ than *Leg* A,B,C,D (figure 34 shows alignment of amino acid sequences of all 115 storage protein genes so far determined). Conversely, sequence from 2 other glycinin CDNA species encoding subunits A_2B1a (Marco et al., 1984) and A1a Bx (Negoro et al., 1985) shows much greater homology with *Leg* A types genes from pea. This indicates a clear division of legumin genes into 2 subfamilies and applies to the *Pisum, Vicia,* and *Glycine* genera. According to Baumlein et al., 1986, genes more homologous to the pea 'major' legumin genes (*Leg* A class) may be designated A type, whereas genes more homologous to the pea 'minor' legumin genes (*Leg* J class) may be designated B type. Further discussion on origins of these subfamilies is presented in section 4.2.

The two soybean B-type legumin cDNA sequences are 82% homologous to each other (Momma et al., 1985) but the level is much lower when comparison is with *Leg* J, K and *VfaLeB*$_4$ (approx 60%; data not presented: homologies to all three of these genes are not significantly different.

Once again comparisons show almost complete divergence of sequence at the C-terminal region of the legumin α subunit, agreeing with observations in section 4.1.2.1 that this region varies greatly amongst the legumin genes, the only constraint being a shared richness of A-G nucleotide composition, with encoded amino acids being mainly polar and hydrophilic. Table 8 gives full details of homology comparisons.

3' flanking sequences of the soybean cDNA species show significant homology to those of *Leg* J and K (40 - 51%; Table 8). Strongest conservation is at the 3 putative polyadenylation sites, but is also evident in several relatively G-C rich regions (e.g. bases 1838 - 1848, 1924 - 1932, 1972 - 1979 in *Leg* J), possibly suggesting some functional importance at these sites. A comparison of the levels of homology between 3' flanking sequences in A- and B- type legumin genes (for pea *Leg* A against soybean A_2B1a, 46% to the limit of homology, compared with 40 - 51% quoted above for pea *Leg* J, K against soybean A_4 B_3 and A_5 A_4 B_3) suggests a similar rate of evolution in this region between A- and B- type legumin genes.

Table 7 Homologies between Leg J, Leg K and the Vicia faba B-type legumin gene LeB4

DNA SEGMENT	LENGTH bp	SUBSTITUTIONS			DELETIONS			HOMOLOGY %	
		B4 → J + K	B4 → J	B4 → K	B4 → J + K	B4 → J	B4 → K	B4 → J	B4 → K
5' FLANKING									
-450[*] → -226	225	-	48	-	-	32	-	64.4	-
-225 → -1	225	-	30	-	-	13	-	80.8	-
5' NON-TRANSLATED[+]	54	-	7	-	-	1	-	85	-
EXON 1	(458 (101	- 4	34 0	- 1	- 0	0 0	- 0	92.6 96	-) 95)
INTRON 1	138	5	2	2	30	35	51	60	37
EXON 2	495	64	4	7	48	21	0	72.3	76.0
INTRON 2	117	6	8	9	24	8	2	59	64
EXON 3	384	10	21	5	0	6	0	90.4	96.1
3' FLANKING	(265[§] (35	16 1	32 28	6 0	1 1	21 0	1 0	73.6 NS	90.9) 94)

[*] No significant homology between Leg J and Vf LeB4 5' to this

[+] Taken to start codon in LeB4 (codon 6 in Leg J)

[§] No significant homology to Leg J 3' to this

Table 8 Homology comparisons of soybean B-type legumin cDNA species (A3B4, A5A4B3) to _leg_ J and _leg_ K

SEQUENCE	TOTAL BASES	SUBSTITUTIONS			DELETIONS			HOMOLOGY %	
		A3B4 → J + K	A3B4 → J	A3B4 → K	A3B4 → J + K	A3B4 → J	A3B4 → K	A3B4 → J	A3B4 → K
EXON 1+	(463		135	–	–	24	–	65.7	–)
	(101	22	0	1	0	0	0	78	77)
EXON 2	681	175	8	13	123	9	12	54.2	52.3
EXON 3	384	105	15	6	27	6	0	60.2	63.5
3' FLANKING	258	42	10	6	88	1	20	45.3	39.5
		A5A4B3 → J + K	A5A4B3 → J	A5A4B3 → K	A5A4B3 → J + K	A5A4B3 → J	A5A4B3 → K	A5A4B3 → J	A5A4B3 → K
EXON 1+	(463	–	130	–	–	24	–	66.7	–)
	(101	21	0	1	0	0	0	79	78)
EXON 2	774	183	7	9	216	9	12	46.4	45.7
EXON 3	399	106	14	6	15	6	0	64.7	67.7
3' FLANKING	247	33	13	7	75	0	21	51.0	44.9

+ FROM CODON 6 IN LEG J

4.2.1 Expression of Legumin genes

4.2.1.1 Leg D

Hybridisation experiments detailed in sections 3.1.3 and 3.1.4. give clear evidence that the *Leg* D gene is not transcribed. Using a probe specific only to the 3' flanking region of *Leg* D no hybridisation to pea messenger RNA was detected, in accordance with the assignment of *Leg* D as a pseudogene. The messenger RNA species of 2000 bases hybridising to the entire *Leg* D (see fig. 4) gene must therefore represent cross-hybridisation with transcripts of *Leg* A. The failure to detect an mRNA species corresponding to *Leg* D does not prove that no transcription of this gene occurs, transcription may occur at a very low level, or may result in a highly unstable RNA species.

4.2.1.2 Leg J and Leg K

The 2300 base band obtained by hybridising a 3' flanking sequence probe from *Leg* K to pea cotyledon RNA corresponds to message transcribed from *Leg* K, hence indicating its status as an expressed gene. The less intense 4200 base band is of unknown origin (see fig. 11).

The results of hybridisation of coding and 3' flanking sequence probes from *Leg* J to messenger RNA suggest that the coding sequence hybridised to mRNA species expressed from all members of the subfamily, giving a heterogeneous range of bands from 2100-2400 bases, whereas the 3' flanking sequence probe was picking up mRNA species specifically expressed from *Leg* J to give a single band of approx. 2100 bases. The increased strength of hybridisation correlating with advance in cotyledon development mimics similar findings with *Leg* A (Gatehouse et al., 1982), in that expression is at a low level in cotyledons in the early stages of development and steadily increases during cotyledon expansion.

Further evidence for the expression of these genes can be taken from the sequences of two cDNA clones, isolated by screening cDNA libraries prepared from poly A+ RNA purified from developing seeds. One cDNA, pLG 3.121. (Gatehouse, 1986) corresponds exactly in its nucleotide sequence to bases 503-1456 in *Leg* J (excluding introns). A second cDNA, pCD40, isolated from pea variety 'Birte' (Domoney and Casey, 1984, corresponds to bases 919 - 1940 in *Leg* K (excluding IVS-2), with only one base substitution between cDNA and genomic clone (a C in the cDNA, and a T in the gene at 1706).

4.1.3 Vicilin Genes

Analysis of the important features of pea vicilin genes was made possible by determination of the complete nucleotide sequence of *Vic* B (see fig. 15) and partial nucleotide sequence of *Vic* C (see fig. 20)

4.1.3.1. *Vic* B

This 50,000 Mr polypeptide encoding gene appeared to contain most of the features characteristic to all eukaryotic genes.

However, of prior importance, in that it may have influenced such features in the gene, is the interruption of the coding sequence at the 3' end of the gene after only 12 amino acid residues of exon 6 by sequence of unknown origin. Indeed, the experiment carried out in section 3.3.4, where no evidence was found to suggest that the missing 3' end of the gene was present in the 6.4 kb of sequence adjacent to the 3' end of the gene, seems to suggest that the gene may no longer to transcribed. Although no evidence exists to substantiate this theory, the fact that the gene no longer possesses a polyadenylation site(s) implies that any message transcribed would be unstable, since absence of a poly (A) tail would render the message susceptible to exonucleolytic attack (Bergmann and Brawerman, 1977). Also absence of a termination codon in any surviving message would potentially interfere with the efficiency of translation. Surprisingly, a similar situation of interruption in 3' coding sequence appears to exist in another pea vicilin gene, *Vic* J, coding for a 47,000 Mr polypeptide (Bown et al., manuscript in prep). Here the interruption occurs at an even earlier point than in *Vic* B, this time at the end of Exon 5, and again no evidence for the existence of the missing 3' end could be found within the next 5.7 kb of sequence.

As stated, the origins of this interrupting sequence are unknown. However, when a fragment specific to this region at the end of *Vic* B was radio labelled and hybridised to pea genomic DNA, the results indicate some degree or repetition of this sequence in the pea genome. These findings do lead to one possible explanation for the origins of this sequence, namely that it may be due to a transposon insertion event, in that transposons can be as long as 10-12 kb, and also that sequences at each end of the transposon <u>near</u> but <u>not</u> including the short inverted repeats, have been found to occur as repeats (of up to 50 copies) in the genome in which the transposon is present (Nevers et al., 1984).

Potentially as a result of this damage to the 3' end of the gene, it was considered a possibility that other important features of the gene may have become altered due to an implied lack of functional constraint. However, the only possible example of this is at the site of greatest homology to the 'CAAT' box, occurring in *Vic* B as the sequence CAAC,77 bp away from the cap site (see below). Indeed, not only is the last T of the element substituted for by a C, but the nucleotides surrounding the box also differ slightly from those associated with a conventional 'CAAT' box; ie. in *Vic* B the sequence runs GGCCAACTGT, whilst the classic consensus sequence is GGC/T CAATTCT. As to whether such an altered sequence would still be functional is not known.

Also present in the 5' flanking sequence is a 'TATA' box, the second T being 30 bp upstream of the cap site. Other potential regulatory elements in this region will be discussed in section 4.1.3.4.

The gene appears to have 5 introns, as has been found for other 7S storage protein genes - both β and α type Phaseolins (Slightom et al., 1983, Slightom et al., 1985) α 'type β -conglycinin (Gma α ' Doyle et al., 1986) and *Vic* J (Bown et al., manuscript in preparation). Relative positioning of these introns between these genes will be discussed in section 4.1.7. Lengths of each of the introns differ between the genes considerably. In *Vic* B, they are as follows, with figures for a phaseolin β -type gene (*Pvu* β Slightom et al., 1983) in brackets as an example for comparison: IVS - 1, 337 bp (72), IVS - 2, 171 bp (88), IVS - 3, 103 bp (124), IVS - 4, 240 bp (128), and IVS - 5, 133 bp (103). In all cases, the introns in *Vic* B conform to the GT/AG boundary rule (Breathnach et al., 1978).

In the coding region, presence of a 23 amino acid residue leader sequence is indicated by comparison of the deduced amino acid sequences obtained by N-terminal sequencing of several mature vicilin polypeptides (Lycett et al., 1983). The amino acid residues either sie of the cleavage site (see fig. 15) do not conform to those identified by Von Heijne (1983) as the most suitable for these positions in eukaryotes. It would seem from a comparison of amino acid sequences around the two potential other regions of post-translational cleavage that *Vic* B encodes a polypeptide designated as Type A (Lycett et al., 1983) - ie no cleavage at either the α - β or β - γ sites, since in both these regions, Ser-Leu-Lys at α - β , and Lys-Glu-Asp at β - γ , the amino acid composition appears sufficiently different from the suggested sequence Lys-Glu-Asn (with Asn appearing to be the crucial residue) (Gatehouse et al., 1982) as to imply that these are non-cleavage sites in *Vic* B. Additionally, no sites with consensus to the potential glycosylation signal N(Asn)-X-T(Tyr)/S(Ser) (Marshall, 1974) were found.

4.1.3.2 *Vic* C

Unfortunately due to time limitations, nucleotide sequencing of *Vic* C was not complete when this thesis was written. However, as can be seen from fig. 20, sequencing did cover approx 0.5 kb of 5' flanking sequence, continuous with coding sequence extending 68 bp into IVS-1, a central section covering all of IVS-2 and IVS-3 and including all of exon 3 and 159 bp of exon 4, another section giving 110 bp at the 3' end of IVS-4 along with 143 bp of exon 5, and finally a section at the 3' end of the gene covering the last 149 bp of exon 6, along with 151 bp of 3' flanking sequence.

Several interesting features emerge from these data on this second gene coding for a 50,000 Mr polypeptide. Firstly, unlike *Vic* B (and indeed *Vic* J), the 3' end of *Vic* C appears to be intact, complete with a termination codon and a polyadenylation signal 59 bp downstream of this. Hence in this respect at least, the gene would appear to be functional. However, a termination codon does seems to exist after 28 amino acid residues of exon 4. Further confirmation of this is required from sequencing of the opposite strand of DNA in this region to be sure this stop codon is present. Should this prove to be the case, then the implication would be that *Vic* C is a pseudogene, presumably having only recently achieved nonfunctional status, since this termination codon appears to be the only indication that the gene may no longer be functional. The only other possible area of the gene under question is the 28 bp stretch before the presumed start of IVS-2. The deduced amino acid sequence in this region doesn't appear to match that observed in the corresponding regions of *Vic* B or *Vic* J in any way. Again though, before any conclusions can be drawn, sequence from the opposite strand in this region is also needed, as is sequence further upstream, which may well indicate that IVS-2 in fact starts further upstream than is at present assumed.

Apart from these findings the rest of the data from the gene indicate the following:

In the 5' flanking region, there is clearly a 'TATA' box present, with the second T 31 bp upstream of the cap site (see below). Comparison of the region of potential 'CAAT' box homology in *Vic* B with the corresponding region in *Vic* C, shows there to be a good degree of sequence conservation. However, the sequence in this region of *Vic* C, namely GGCCAAATG, 49 bp upstream of the 'TATA' box, shows even less agreement to the 'CAAT' box consensus sequence, GGT/CAATCT than does *Vic* B (see section 4.1.3.1). The remarkable conservation of sequence in this area of these genes is discussed further in section 4.1.2.4., along with any other areas of potential regulatory importance. Within the coding region 4 of the 5 introns found in other 7 S storage protein genes are shown to be present (relative positioning will be discussed in section 4.1.7) but with the data only definitely giving complete sequence for IVS-3, 90 bp in length (as compared to 104 bp in *Vic* B and 124 bp in phaseolin - see section 4.1.3.1). Of the intron/exon borders determined, each conforms to the GT/AG boundary rule (Breathnach et al., 1978).

As is the case with *Vic* B (see section 4.2.3.1), *Vic* C appears to code for a Type A polypeptide - no sites resembling the Lys-Glu-Asn amino acid sequence required for cleavage appear to be present, instead the polypeptide has the amino acid sequences Arg-Gly-Leu-Arg at the region corresponding to the α - β site, and Arg-Glu-Glu-Asp at

Arg-Glu-Glu-Asp at the β - γ site. Unfortunately the region corresponding to that identified as a glycosylation site in 50,000 Mr Vicilin (Lycett at al., 1983) is not present in the data so far obtained for *Vic* C.

4.1.3.3 Hybridisation of *Vic* B and *Vic* C to Genomic DNA

Hybridisation of *Vic* B to a Southern blot containing restriction digests of genomic DNA (see section 3.3.5 and fig. 16) showed strong hybridisation to the corresponding 4.7 kb Eco R1 fragment. Unfortunately, gene copy equivalents on this Southern blot failed to hybridise properly, consequently the number of gene copies in this band could not be deduced. However, the result does seem to conform to the pattern shown when the cDNA pCD48 (see section 3.3.1) was hybridised to Eco R1 digests of genomic DNA from 6 different pea genotypes (Domoney and Casey, 1985). Here a band representing 3 - 4 gene copies was detected at approx 5kb, along with a feinter band, at the 1 or 2 copy level, at 5.5 kb. Such a secondary feinter band was also detected at 5.5 kb by *Vic* B in the Eco R1 digested genomic DNA. As well as the 4.7 Kb Eco R1 band, one of almost equal intensity was detected by *Vic* B at 2.3 kb in the Hind III digest, corresponding to the equivalent fragment present within the Eco R1 fragment of pJC1-16 (see fig 14), thus representing most of the Vic B coding sequence. Other bands detected in the other digests appear relatively weak in comparison, but since most are greater than 9 kb this might be explained by incomplete transfer of the larger DNA fragments during transfer from agarose gel to Nitrocellulose filter.

Hybridisation of *Vic* C to an equivalent Southern blot resulted in an altogether more complicated pattern of bands on the autoradiograph (see section 3.4.3 and fig. 21). Again, for an unknown reason, the gene copy equivalents failed to be detected. Of those bands detected in the genomic digests it was surprising to find so many in the Eco R1 track. A tentative assignment of 1 gene copy might be made for the 4.1 kb *Vic* C equivalent fragment, with bands of stronger intensities (approx 2 - 4 copies?) at 4.7, 5.0, 6.9, 7.8 and 8.9 kb. The band at 7.8 kb is the strongest of all, and possibly represents cross hybridisation to another set of vicilin genes yet to be isolated. However, it might also be due to cross hybridisation with a set of *Vic* J type genes. *Vic* J has been shown to exist within a 7.4 kb Eco R1 fragment (Bown et al., manuscript in press). Such a strength of cross hybridisation would not have been expected though. Patterns of hybridisation in the other digests do little to shed more light on the gene copy number situation - one point possibly worth noting though is the two relatively strong, large bands (approx 25 and 17.5 kb) in the Bam H1 track. These would seem to indicate high levels of copy number, since even though transfer from the agarose gel to the nitrocellulose filter may have been incomplete, the intensity of these bands is still strong relative to levels of hybridisation in the other digests.

Overall, the findings from these two experiments suggest that further experiments may be necessary to confirm the published total gene copy number for pea vicilin genes of 11 (Domoney and Casey 1985), with these results suggesting only limited support.

4.1.3.4 Other Potential Regulatory Elements in the 5' Flanking Regions of Vicilin Genes

Apart from the features of the 5' flanking sequences from *Vic* B and *Vic* C which have already been discussed, comparison of these sequences with corresponding regions of *Vic* J (Bown et al., manuscript in preparation) a convicilin gene from pea, CVA (Bown et al., 1988), *Pvu* β from *Phaseolus vulgaris* (Slighton et al., 1983),and the *Gma* α ' gene from *Glycine max* (Doyle et al., 1986), indicate the presence of some other potential regulatory elements. Such comparisons were performed in 3 ways: scanning by eye, use of Dot Matrix comparison (using a BBC microcomputer), and with the nucleic acid sequence alignment program NUCALN (on an IBM personal computer)

Firstly, these analyses confirmed the presence of a conserved sequence of 42 bp previously termed the 'Vicilin box' (Gatehouse et al., 1986). It appears to be present in all 6 of the genes studied, at a position ranging from 88 to 106 bp relative to the 'TATA' box (see fig 29). As can be seen from this, the sequence can conveniently be divided into two regions; a 5' region of 13 bases, where all 6 genes match in 12 of the 13 positions, and which also includes a sequence identified in *Leg* A as being homologous to the adenovirus core enhancer element (Lycett et al., 1984) - also shown in fig. 29. The second, less conserved, 3' region shows no homology with the legumin box (Baumlein et al., 1986). Overall this 'vicilin box' appears to be a good example of a conserved upstream sequence specific to the 7S storage protein gene family. As such, it is possibly analogous to the tissue-specific transcriptional enhancer sequences found in similar positions in other eukaryotic genes (another possible example of this in plant genes has been identified by Kreis et al., 1986 in barley prolamin storage protein genes).

Another important conserved sequence found was that at the cap site, or transcription start (see section 1.1.2.1) No S1 mapping experiments (Berk and Sharp, 1977) were performed on *Vic* B or *Vic* C to determine this site, but results from such work done on the *Vic* J(Bown et al., manuscript in preparation) and CVA (Bown et al., 1988) genes and also *Pvu* β (Slightom et al., 1983) and *Gma* α ' (Doyle et al., 1986) showed initiation of transcription to occur at the sequence CATC in each case. Such a sequence was also observed in both *Vic* B and *Vic* C, at positions corresponding to those found in the other genes (see fig. 30). Although this region would not possess enhancer - like properties, it is interesting to note the same sequence being involved with transcription initiation in all these 7S genes, as well as in the case of *Leg* A (Lycett et al., 1986). Such a finding indicates that a similar mechanism of initation and capping is used in all storage protein genes.

Finally, the 5' flanking sequences were scanned for sequences homologous to the Mammalian virus enhancer element (Weiher et al., 1983, Gruss, 1984) (with several already having been found for *Pvu* β and *Gma* α (Doyle et al., 1986), and also to the Adenovirus core enhancer element (Hearing and Schenk 1983). Since enhancers can work in either orientation with respect to a gene (see section 1.1.2.1) both the sequenced strand and its deduced complementary strand were

Fig. 29 The putative 'vicilin box'. The conserved sequence region in the 5' flanking sequences of 3 pea vicilin genes. 1 pea convicilin gene (CVA), 1 phaseolin gene (*Pvu* β) and 1 soybean β conglycinin gene (*Gma* α) plus pea *Leg* A is boxed (see text for references). The 'legumin box' in *Leg* A is indicated by broken lines. The sequence in *Leg* A identified as being homologous to the adenovirus enhancer core element is underlined.

tative "Vicilin Box"

```
Vic B          TTCATTTTGCCACTTCAATTTTTGTACATTTTAACACACGTCCATATGCATGGCACAAC  *=100bp to "TATA"
                       *
Vic C          TTCGCTTTGCCACCTCAATTTTTGTACATTTCAACACACGTCCATATGCATGGCACAAC  *= 99bp to "TATA"
                       *
Vic J          TTTAATTAGCCACCTCAATTTTTGTTCATTTCTACACTAGTCAACATGCATGGCAAACT  *= 98bp to "TATA"
                       *
CVA            ACTCAGTTGCCACCTCTATTTTTGTTCATTTCAACACTCGTCAAGTTACATGACACAAT  *= 99bp to "TATA"
                       *
Pvu b          TCTCTTCCGCCACCTCAATTT-CTTCACTTCAACACACGTCAACCTGCATATGCGTGT  *= 95bp to "TATA"(1),106bp to "TATA"(2)
                       *
Gma a          TCTCTTCCGCCACCTCATTTTTTGTTTATTTCAACACCCGTCAAACTGCATCCCACCCC  *= 88bp to "TATA"
                       *
Leg A          AGCCATTAGCCACCTCCTCTATCAGACATAGGTGTAAAGCATTATGCTTCCATAGCCA  *= 125bp to "TATA"
                       *
```

scanned, with the findings summarised in fig.30, showing that several sequences of differing levels of homology to these two viral enhancers were found. The positions of some of these sequences are found much further upstream in the 5' flanking sequence than is the case for the functional virus enhancers (generally within 100 bp upstream from the mRNA cap site - Gruss, 1984). Indeed, experiments involving deletion mutants of an α ' type β -conglycinin gene followed by assessment of levels of expression of the gene in transgenic petunia plants, showed that presence of the region containing the exact match to the Mammalian viral enhancer (GTGGATAG, 611 bp upstream of ATG, see fig 30) appeared to cause no enhancement of expression relative to the level obtained from the gene where this area of sequence had been deleted (Chen et al., 1986). This seems to suggest that such sequences displaying homology to viral enhancers do not appear to exert a similar effect on the plant gene concerned. It must also be said to contradict with findings in other plant genes, where sequences homologous to GTGGA/TA/TA/TG in the 5' flanking region have been shown to have an important enhancing effect on expression of these genes in transgenic plants (Timko et al., 1985, Kaulen et al., 1986, Odell et al., 1985).

Fig. 30 Other potential regulatory sequence elements and consensus sequences in 7S storage protein genes. Sequences represented; *Vic* B, C and J, vicilin genes from pea, CVA a convicilin gene from pea, *Pvu* β a phaseolin gene from *Phascolus vulgaris* , *Gma* α 'a β conglycinin a subunit gene from soybean, and *Leg* A the Legumin A gene from pea (see text for references)

1. Mammalian Virus Enhancer Consensus Sequences(Gruss,1984.Weiher et al.,1983)

Viral Consensus:	G T G G T T T G
	A A A

Gmaa'	G T G G A T A G(-611)*
	C A T C C C A C(-145,opposite orientation)
Pvub	G T G G A C T A(-581)
	T T G G T A T G(-539)
	G T G G G T T G(-509)
	C A A C C C A C(-283,opposite orientation)
	C C A A C C A C(-131, ")
VicB	G T T G T A G G(-347)
	G G T G C A T G(-120)
VicC	G G T G A G A G(-355)
VicJ	G G T G C A T G(-116)
CVA	C A A A C T C C(-283,opposite orientation)
	C C A T C C A C(-103, ")

(Consensus sequencse for Gmaa' and Pvub taken from Doyle et al.,1986)

2. Adenovirus Core Enhancer Consensus Sequences (Hearing and Schenk,1983)

Viral Consensus:	A G G A A G T G A C A

N.B.All the following are in opposite orientation

VicB	G C C A C T T C A A(-183,70% homologous)
VicC	G C C A C C T C A A(-171,60% ")
VicJ	G C C A C C T C A A(-178,60% ")
CVA	G C C A C C T C T A(-165,60% ")
Pvub	G C C A C C T C A A(-216,60% ")
Gmaa'	G C C A C C T C A T(-178,70% ")
(q.v.LegA	G C C A C C T C C T(-196,80% "))

3. Transcription Start Sequence

VicB	C A T C(-54)
VicC	C A T C(-54)
VicJ	C A T C(-48)
CVA	C A T C(-35)
Pvub	C A T C(-78)
Gmaa'	C A T C(-56)
(q.v.LegA	C A T C(-34))

*=Position relative to ATG

Whether or not these sequences homologous to viral enhancers play a functional role in the 7S storage protein genes, it is interesting to note that the region homologous with the adenvirus core enhancer lies at the 5' end of the potential 'vicilin box' of each of the 6 genes. It remains open to speculation as to whether this finding is of any significance, but one possible theory might be that the viral enhancer may have been inserted originally into the 5' flanking sequence of a long distant ancestral gene, possibly giving a degree of enhancement of expression. Subsequent evolution of the sequence at this region may then have led to the enhancer developing a tissue specific role, still present in these 7S genes due to conservation of this sequence during the divergence of the genes during evolution. Such a pathway would not seem appropriate for the evolution of the 'legumin' box in the 11S storage protein genes - firstly because the region of homology to the adenovirus enhancer in *Leg* A is 32 bp from the legumin box, and secondly that such a region of homology to the viral enhancer does not seem to be present in the B type *Vicia faba* Legumin gene (*Vfa Le B* $_4$ Baumlein et al., 1986).

4.1.3.5 Comparison of the nucleic acid sequences of *Vic* B and *Vic* C with those of other 7S storage protein genes.

By once again using the nucleic acid sequence alignment programme NUCALN, it was possible to compare separately the 5' and 3' flanking sequences, each exon, and each intron of *Vic* B and *Vic* C against the corresponding regions of *Vic* J, CVA, *Pvu* β and *Gma* α '. Each individual comparison was analysed to give a percentage level of

TABLE 9 Nucleic acid sequence comparisons of VicB and VicC against other 7S storage protein genes.

	VICB	VICJ	CVA	PHAS(Pvu b)	CGLY(Gma a´)
VICC 5'FL	1st 278 82.4	1st 234 70.9	1st 186 53.4	1st 258 57.4	?
VICB 5'FL	-	1st 299 67.2	1st 184 73.9	1st 175 56.0	?
VICC 5'FL	next 247 47.8	next 225 47.6	next 313 46.6	next 254 32.3	?
VICB 5'FL	-	next 200 41.7	next 314 66.6	next 400 43.8	?
VICC EX1	84.4	82.3	72.0	67.5	65.4
VICB EX1	-	78.8	70.6	64.0	67.1
VICC EX2	-	-	-	-	-
VICB EX2	-	82.9	74.4	66.4	63.6
VICC EX3	88.9	85.2	76.0	65.4	65.4
VICB EX3	-	82.7	77.3	69.1	69.1
VICC EX4	86.7	80.4	67.7	46.8	59.7
VICB EX4	-	86.3	72.6	60.6	65.6
VICC EX5	89.4	85.7	73.9	58.4	59.1
VICB EX5	-	85.1	76.3	72.3	72.1
VICC EX6	82.0	-	72.5	55.8	67.8
VICB EX6*	-	-	71.4	45.2	63.7
VICC 3'FL	-	-	63.6**	57.2	52.6
VICB 3'FL	-	-	-	-	-
VICC INT1	60.6	-	61.2	-	-
VICB INT1	-	-	47.3	-	-
VICC INT2	61.6	-	46.8	-	-
VICB INT2	-	-	70.9	-	-
VICC INT3	67.0	-	40.8	-	-
VICB INT3	-	-	46.9	-	-
VICC INT4	51.3	-	50.7	-	-
VICB INT4	-	-	48.9	-	-
VICC INT5	-	-	-	-	-
VICB INT5	-	-	50.0	-	-

* Taken from pDUB9,the equivalent cDNA to VicB
** Figure obtained from the first 77bp of 3' flanking sequence of each gene.A low level of homology exists past here
***Upstream from ATG

homology (where gaps occurred in the alignment, to maximise base matching, these were excluded from the calculations). The results of these calculations are given in Table 9. Unfortunately, for some unknown reason, the programme failed to produce alignments between the 5' flanking sequences of *Vic* B and *Vic* C against *Gma* α '. As far as other comparisons in this 5' region were concerned, it was deemed appropriate to divide the alignment into two parts, since in each case a fairly distinct dividing line existed between two different levels of homology, thus giving the two figures shown for each comparisons in Table 9. Had time permitted, the alignments would have been further analysed to give figures for silent and replacement substitutions in regions of coding sequence comparison. Also, intron comparisons were limited to those between *Vic* C and *Vic* B, *Vic* C and CVA and *Vic* B and CVA.

The results reveal several points of interest:

i) *Vic* C vs *Vic* B: Overall, a very high level of homology between these two genes can be observed, with levels around or above 85% for all the exons, above 80% for the first 278 bp of 5' flanking sequence and nearly 50% for the remaining 247, and even an appreciable level still existing between the introns - as high as 67% in intron 3. A more rigorous assessment of relatedness between all the genes will be made in section 4.2.2., but these figures clearly indicate that these two genes both coding for 50,000 Mr polypeptides, have diverged relatively recently in evolutionary history.

ii) *Vic* C and *Vic* B vs *Vic* J: This represents a comparison of two genes coding for 50,000 Mr polypeptides, with one coding for one of

47,000. Hence it is interesting to note that overall, *Vic* C appears to be more homologous to *Vic* J than does *Vic* B. Of the regions compared, only exon 4 shows *Vic* B to have the higher level of homology - other than this, levels observed for *Vic* C are consistently slightly higher. The differences in levels observed do not appear to be sufficiently high to contradict the theory that the 50,000 and 47,000 Mr genes arose by a duplication of a common ancestor. Following this, evolutionary modification would cause the genes to diverge, with the two observed 50,000 Mr genes then resulting from a further duplication event, and as a consequence displaying similar (but not identical) levels of homology to the 47,000 Mr gene.

iii) *Vic* C and *Vic* B vs CVA: A similar situation to that observed in the comparison with *Vic* J seems to exist with CVA, especially in the exon comparisons, suggesting an analogous pattern of gene evolution to that suggested above with *Vic* J, ie duplication from a common ancestor, subsequent modification of one of the duplicates (possibly by some insertion event - Bown et al., manuscript in press) to produce the much larger convicilin gene, and then later duplication to produce two 50,000 Mr genes.

It is interesting to note however the relatively large differences in levels of homology between vicilin genes and CVA shown in the 5' flanking sequence - 53.6% for *Vic* C over the first 186 bp as against 73.9% for *Vic* B over the first 184 bp, and 46.6% for *Vic* C in the remaining 313 bp, but 66.6% for *Vic* B on the remaining 314 bp. The reason for such a difference in levels of sequence conservation in this region is unknown. It is also interesting to note the different levels of homology observed in the first two introns - *Vic* C vs CVA shows 61.2% in IVS-1 compared to 47.3% here for *Vic* B vs CVA, while *Vic* B vs CVA shows 70.9% in IVS-2 as against only 46.8% for *Vic* C vs CVA. This seems to suggest that the intron sequences are relatively free of sequence constraint - also partly borne out by the generally low levels observed with both genes in the other introns (generally around 40-50%). Analysis later in this discussion (see Section 4.1.7.2) showing that the intron positions remain highly conserved in all 7S storage protein genes tends to bear out the principle that it is intron position, rather than sequence, which is important for gene function (Shah et al 1983). Also any homology shown between introns tended in general to be located near the 3' and 5' splice sites, rather than the central region, implying that conservation of sequence is most important in these areas (see also Section 4.1.6).

iv) *Vic* C and *Vic* B vs *Pvu* β : Here the levels of homology are broadly similar across the 5' flanking regions and exons 1 and 3. However notable differences in level can be observed between *Vic* C and *Vic* B over exon 4 and 5, where *Vic* B is more homologous to *Pvu* β , and in exon 6, where the reverse is true. A rather unsatisfactory explanation for this might just be that these (rather pronounced) differences are simply a consequence of the particular pathways of evolution embarked on by the genes after their presumed origins from a common ancestor. As partial support for this, the levels of homology shown in the data are sufficiently high as to agree with the previously suggested concept of a common ancestral origin for these 7S storage protein genes (Borrotto and Dure, 1987). Also, here one can observe a marked general drop in levels of homology compared with those observed when *Vic* C and *Vic* B are compared with genes in the same species, implying that speciation between pea and french bean occurred before the duplication of these genes in pea.

v) *Vic* C and *Vic* B vs *Gma* α ': In this case, only the exon data was available for comparison. Generally, overall homology levels are comparable with those achieved with *Pvu* β (indeed both *Vic* Cand *Vic* B show exactly the same levels in exon 3). One apparent difference between *Gma* α 'and *Pvu* β is in exon 6, the former showing an increase in homology here with *Vic* C and an even greater increase with *Vic* B - the only implication that can be taken from this is that some sort of functional constraint common to pea and soybean operates on exon 6 (possibly in terms of common packaging or degradation processes) which is not present in french bean.

The otherwise broad similarity in levels of homology between *Gma* α 'and *Pvu* β again supports the theory of a common ancestral gene, and possibly even suggests that the 7S genes of soybean and french bean evolved on a separate pathway right after the first duplication from the ancestral gene (see section 4.2.2 for further discussion)

4.1.5 Sequences occurring around the translation initiation

It is now a fairly well established fact, due to several extensive surveys on this subject, that the sequence of bases occurring around initiator codons in eukaryotic genes is non random (eg Kozak 1981, Kozak 1984, Heidecker and Messing 1986, Cavener 1987).

As a result of this work, it became clear that consensus sequences existed in this region, and that thse sequences were seen to differ between plants and animals (Lutcke et al., 1987):

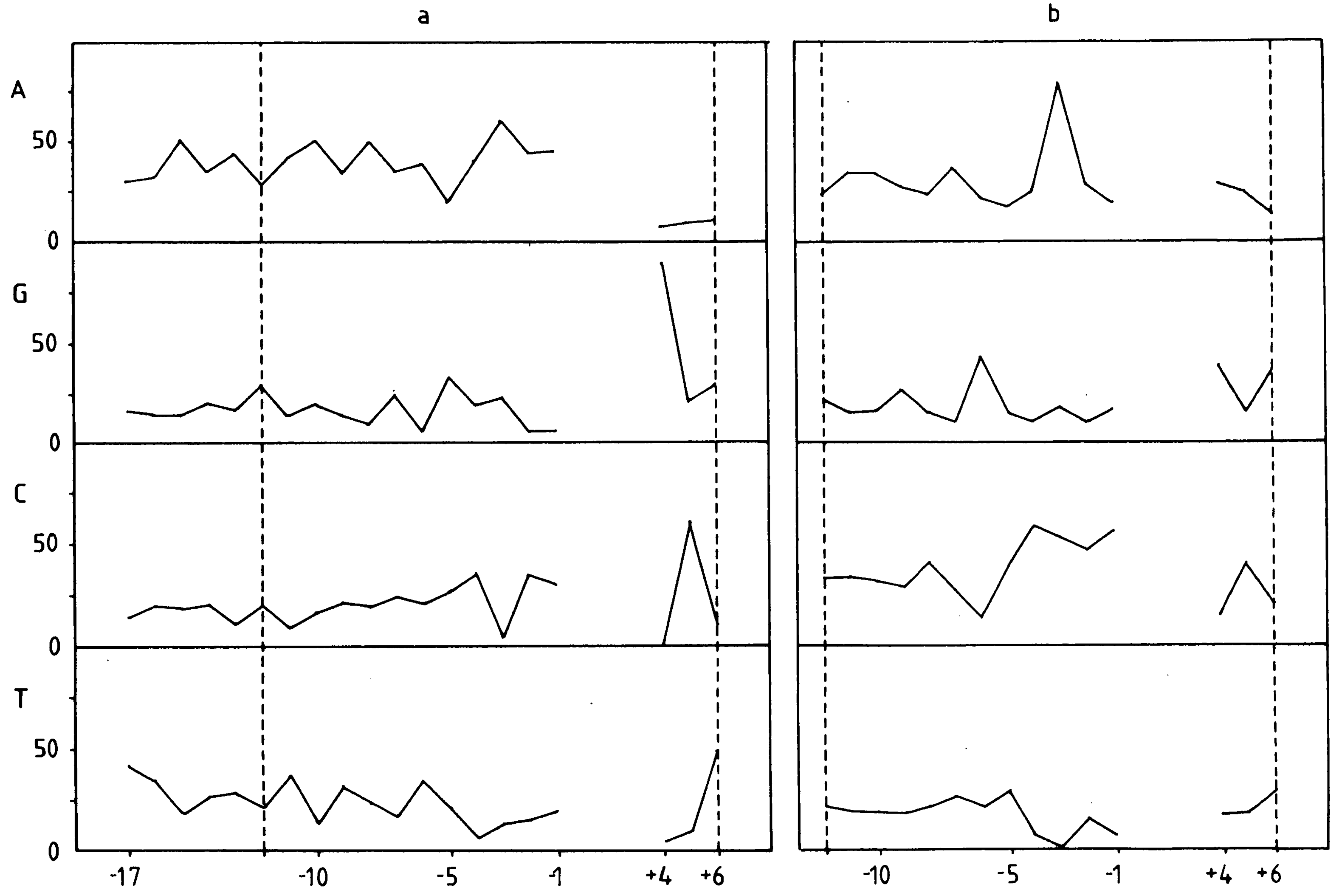
	-4	-3	-2	-1	+1	+2	+3	+4	+5	+6
Animal consensus	C	A	C	C	A	T	G			
Plant consensus	A	A	C	A	A	T	G	G	C	

Comparison over a wider span of sequences can be seen in fig 31. Here percentage frequencies for each of the 4 bases were plotted for positions -17 to +6 (with the A of ATG being +1) for plant genes, and -12 to +6 for animal genes. The data for animal genes was obtained from the sequences in the paper by Kozak, 1984. Plant gene data was combined to give a greater sample size, from two sources - firstly the figures presented by Heidecker and Messing, 1986, obtained from 47 plant nuclear genes, together with the data, compiled by myself, from a further 19 plant genes (references for these genes were obtained from Brown 1986 - all gene references were checked, and any containing sequence data for the intiative codon were used, and were as follows 1) From soybean - Leghaemoglobin *Lb* (Brisson and Verma, 1982), *Lbc* 2 and *Lbs* 3 (Wiborg et al., 1982), plus Nodulin 23 (Mauro et al., 1985) and Nodulin 24 (Katinakis and Verma 1985) ii) From maize - Sucrose synthetase (Werr et al., 1985), *WX* + (Klosgen et al., 1986), *hsp* (heat shock protein) 70 (Rochester et al., 1986) and Triose phosphate isomerase (Marchionni and Gilbert, 1986). iii) From potato - Patatins *pg* T5 (Rosahl et al., 1986) *Sb* 6B and *SA* 10C (Pikard et al., 1986) and Proteinase inhibitor II (Keil et al., 1986) iv) From carrot - Extensin (Chen and Varner, 1985) v) From tobacco - ATP synthase

(Boutry and Chua, 1985) vi) From alfalfa - Glutamine synthetase (Tischer et al., 1986) vii) From antirrhinum - Chalcone synthase (Sommer and Saedler, 1986) viii) From chlamydamonas - Rubisco small subunit *rbcs* 1 and *rbcs* 2 (Goldschmidt-Clermont and Rahire 1986). As can be seen from the two profiles, outside the consensus regions given above, little difference appears to exist between plant and animal genes, apart from a slightly greater prevalence of C's in positions -12 to -5 of the animal genes. Within the confines of the consensus sequences, it has been shown that whilst animal genes show a marked preference for A at -3, this is not quite as pronounced in plant genes, whilst in plant genes the prevalence of G at +4 is not reflected at all in animal genes. Since a mechanism of interaction between the consensus sequence and a homologous region on 18S rRNA has been suggested (Sargan et al., 1982), any differences shown between plant and animal in this region suggest slightly different sequences in the corresponding regions of plant and animal 18S rRNAs.

Fig 31. Percentage frequency distribution of each nucleotide around the functional initiator codon in a large number of plant and animal mRNAs (for sources, see text). The nucleotide immediately prior to the ATG codon is number -1, nucleotides +4 to +6 represent the start of the protein coding sequence. Dotted lines delineate regions comparable by nucleotide position.

a = plant
b = animal


a
b
A
G
C
T
50
0
-17
-10
-5
-1
+4
+6

In order to assess the situation in the pea seed storage protein genes, sequences available for this region were compiled as shown in Table 10, with the number of bases occurring in each position recorded for comparison against the plant consensus sequence, and also the most frequent base observed from the plant gene data collated as above. The most interesting point to emerge from this data is the apparent lack of preference shown by the pea genes for having an A residue at -3, where only 5/9 genes have As, the others all having Ts. Also interesting are positions -4 (C = 4/9, A only 3/9), -2 (T = 5/9, C only 3/9) amd -1 (C = 4/9, A only 2/9). Although these figures only come from a small sample size, it is nevertheless interesting to note that overall the pea genes seem to show a greater similarity to the animal consensus than to the plant. However, since it has been stated that the most important position in the pea consensus is at +4, where G is greatly favoured, and that having an A at -3 is of lesser importance than in animal genes (Lutcke et al., 1987), then by showing 6/9 G's at +4, the pea genes conform to the plant consensus. As to whether differences observed in the other positions would be significant in inhibiting translation initiation, nothing can be said for sure - data would be needed on the translational efficiencies of genes differing in sequence at each one of these positions. Any sort of evidence of this kind could in fact prove valuable in future work involving efficient translation of foreign genes, perhaps altered in this region during insertion into a vector, in transgenic plants.

TABLE 10 Compilation of sequences around the initiator codon in pea seed storage protein genes.

	-17							-10										+1					+6
LegA	T	C	T	T	A	G	T	A	T	C	T	C	T	C	T	T	C	A	T	G	G	C	T
LegB	T	C	T	T	A	G	T	A	T	C	T	C	T	C	T	T	C	A	T	G	G	C	T
LegC	T	C	T	T	A	G	T	A	T	C	T	C	T	C	T	T	C	A	T	G	G	C	T
LegD	-	T	A	T	A	T	T	C	T	A	T	C	C	A	A	C	T	A	T	G	G	C	T
LegJ	T	C	T	C	T	C	C	T	T	A	G	T	A	G	T	T	T	A	T	G	A	T	C
VicB	A	A	C	A	G	G	C	T	G	C	T	A	C	T	A	C	A	A	T	G	A	A	A
VicC	A	C	A	T	G	G	C	T	G	C	T	A	C	A	A	C	A	A	T	G	A	A	A
VicJ	C	A	T	T	A	A	T	C	C	A	A	A	T	C	A	A	T	A	T	G	G	C	T
CvA	C	T	A	G	T	G	A	A	A	T	A	C	A	A	A	T	C	A	T	G	G	C	G
Totals																							
A	2	2	3	1	5	1	1	4	1	3	2	3	2	3	5	1	2	9	-	-	3	2	2
G	-	-	-	1	2	6	-	-	2	-	1	-	-	1	-	-	-	-	-	9	6	-	1
C	2	5	1	1	-	1	3	2	1	5	-	5	3	4	-	3	4	-	-	-	-	6	1
T	4	2	5	6	2	1	5	3	5	1	6	1	4	1	4	5	3	-	9	-	-	1	5
Plant-Consensus	N	N	N	N	A	N	N	N	N	A	N	N	N	N	A	N	N	A	T	G	G	C	T
Plant-Most Frequent	T	Y	A	A	A	G	A	A	T	A	X	A	N	A	A	C	A	A	T	G	G	C	T
														C									

4.1.6 Intron Consensus Sequences

The interruption of protein coding sequence by introns in a large variety of eukaryotic genes dictates that these introns must be precisely removed from initial mRNA precursors (pre-mRNA or heteronuclear RNA), before the message can be efficiently translated. The process by which this excision takes place is known as splicing.

It has now been established that in order to achieve efficient splicing, several important features must be conformed to by the intron and its bordering exon sequences. After the initial establishment of the GT/AG boundary rule (Breathnach et al, 1978), subsequent more extensive consensus sequences relating to the 5' and 3' boundaries of introns have been derived (Mount, 1982, Brown 1986).

Following this, a suggestion for a general mechanism of splicing was made, whereby introns are removed in the form of a 'lariat' RNA, where the 5' end of the intron forms a 5' - 2' phosphodiester bond with the 2' -OH of a conserved region containing an adenosine residue (the branch point), occurring between 18 to 40 nt upstream of the 3' splice site (Ruskin et al., 1984., Zeitlin and Efstratiadis 1984., Reed and Maniatis, 1985). This concept has now been expanded upon with the elucidation of the order of events involved in pre-mRNA processing (Gerke and Steitz, 1986); Firstly, cleavage occurs at the 5' splice site of the intron, which in turn is followed by the formation of a branched structure in which the G at the 5' terminus of the intron forms the phosphodiester bond with the A in the branch point. Next, excision of the lariat structure occurs by cleavage at the 3' splice site, and finally the exons are ligated together.

It is now also known that the splicing events are dependent on the abundant U-type small nuclear RNAs (U-sn RNAs) present in the form of ribonucleoprotein complexes (U-sn RNPs) (Frendewey and Keller, 1985. Grabowski et al., 1985). Indeed, there appears to be increasing evidence that all species of U-snRNP - U1, U2, U5 and U4/U6 (which exist as a single complex - Bringmann et al., 1984, Hashimoto and Steitz, 1984) are involved in the various stages of processing (Lerner et al., 1980., Black and Steitz 1986), possibly in the form of a so-called 'spliceosome' complex (Grabowski et al., 1985) apparently containing at least 3 of the U-sn RNPs (Sharp, 1987). The importance of these U-sn RNPs is underlined by two observations - firstly that snRNP's from fungi, the yeast *Saccharomyces cerevisiae* and also from peas, can be immunoprecipitated by antibodies raised against human U-snRNPs (Tollervey and Mattaj, 1987), implying a considerable degree of evolutionary conservation. Secondly, the elucidation of the entire nucleotide sequence of broad bean U2 RNA (Kiss et al., 1987), shows that its secondary structure matches that of a model proposed to accommodate all other fully sequenced higher eukaryotic U2 RNAs (Reddy, 1985), again implying considerable functional constraint has operated during evolution in order to allow maintainance of an active role in the splicing process.

Some evidence also exists for the individual roles played by the U-snRNPs within the spliceosome complex. Rat U2 - sn RNA has been shown to possess a single stranded region with sequences homologous to both the branch point and the 3' splice site consensus sequences in mammalian introns (Reddy and Busch, 1983, Keller and Noon 1984). Also, it appears that U1-sn RNA utilizes specific base-pairing in its recognition of the 5' splice site (Zhuang and Weiner, 1986). Considerably more evidence exists to substantiate these and other individual roles of U-sn RNPs but will not be discussed further here.

In view of the fact that a considerable degree of conservation has operated on the U-sn RNA sequences, a comparison between plant and animal intron consensus sequences (Brown 1986) not surprisingly shows a considerable degree of similarity between the corresponding regions:-

i) 5' splice site -

	-3	-2	-1		+1	+2	+3	+4	+5	+6	
Plant	C/A	A	G	:	G	T	A	A	G	T	(: = splice site)

	-3	-2	-1		+1	+2	+3	+4	+5	+6
Animal	C/A	A	G	:	G	T	A/G	A	G	T

- Here we can see the sequences are virtually identical.

ii) 3' Splice site -

						-10						-4			-1		+1
Plant	T	T	T	T Pu	T	T	T Pu	T Pu	T Pu	T Pu	T	$\underline{G}$	C	A	G	:	G

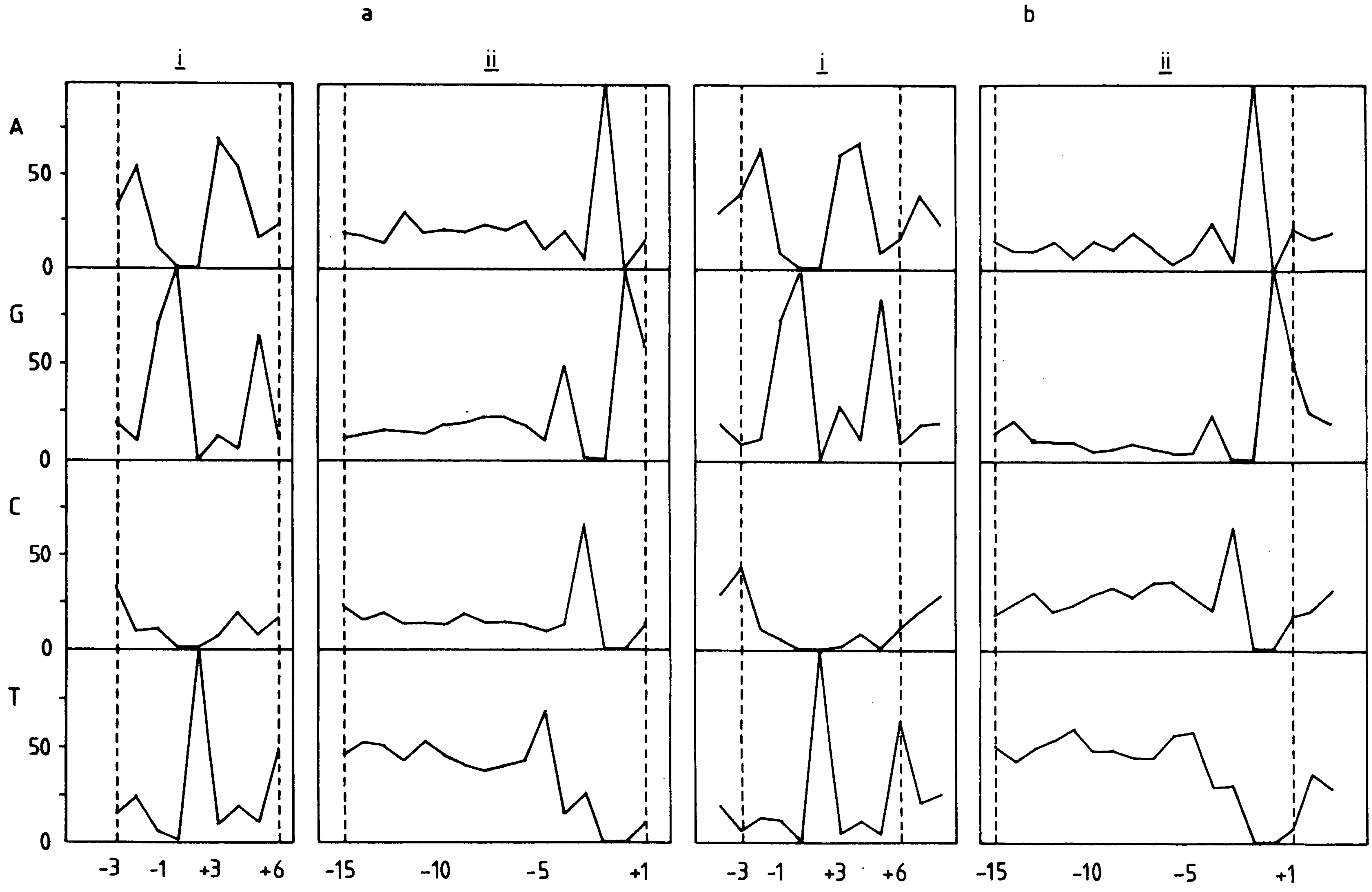
						-10						-4			-1		+1
Animal	T	T	T	T	T	T	T	T	T	T	T	$\underline{N}$	C	A	G	:	G

- Here the emphasis on a poly (T) stretch, present in the animal consensus, is less pronounced in the plant counterpart, and also significant is the difference at position -$\underline{4}$.

The percentage composition of bases at each position for both plant and animal 5' and 3' splice sites is shown in fig. 32 (figs taken from Brown, 1986) illustrating clearly the points made above - generally very similar profiles are observed for both at the 5' site, whilst at the 3' site animal genes show a high level of Ts consistently from -15 to -5, whilst in plants, although Ts tend to predominate in all these positions, A and G also occur in many positions at significant levels. The difference at position -4 is also clear in the 3' site, with plant genes having a 50% presence of G at this site, whilst animal genes show approx 25% of each base here.

Fig. 32 Percentage frequency distribution of each nucleotide around the 5' and 3' intron/exon boundaries of plant and animal genes (for sources, see text). At the 5' (exon/intron) boundary, the GT consensus nucleotides are numbered +1 and +2 respectively. At the 3' (intron/exon) boundary the AG consensus nucleotides are numbered -2 and -1 respectively. Dotted lines delineate regions comparable by nucleotide position.

a = plant
b = animal
i = 5'
ii = 3'


a
b
i
ii
i
ii
A
G
C
T
50
0
50
0
50
0
50
0
-3
-1
+3
+6
-15
-10
-5
+1
-3
-1
+3
+6
-15
-10
-5
+1

iii) The branch point -

	-3	-2	-1	0	+1
Plant	C	T	Pu	A	Py
Animal	C	T	Pu	A	Py

These sequences were taken from the data for plant genes (Brown, 1986) and animal genes (Keller and Noon, 1984) and again is represented graphically in fig. 33 by plotting the percentage occurrence of each base for each position (positions are numbered relative to the branch point nucleotide itself at 0). It is interesting to note that the profiles that result are virtually identical for both plant and animal genes.

Fig. 33 Percentage frequency distribution of each nucleotide around the intron branch points of several animal and plant genes (see text for sources of sequence data). The consensus A nucleotide of the branch point (see text for ref.) is numbered 0, with position +1 indicating the next nucleotide downstream in the 3' direction within the intron. Dotted lines delineate regions comparable by nucleotide position.

a = plant
b = animal

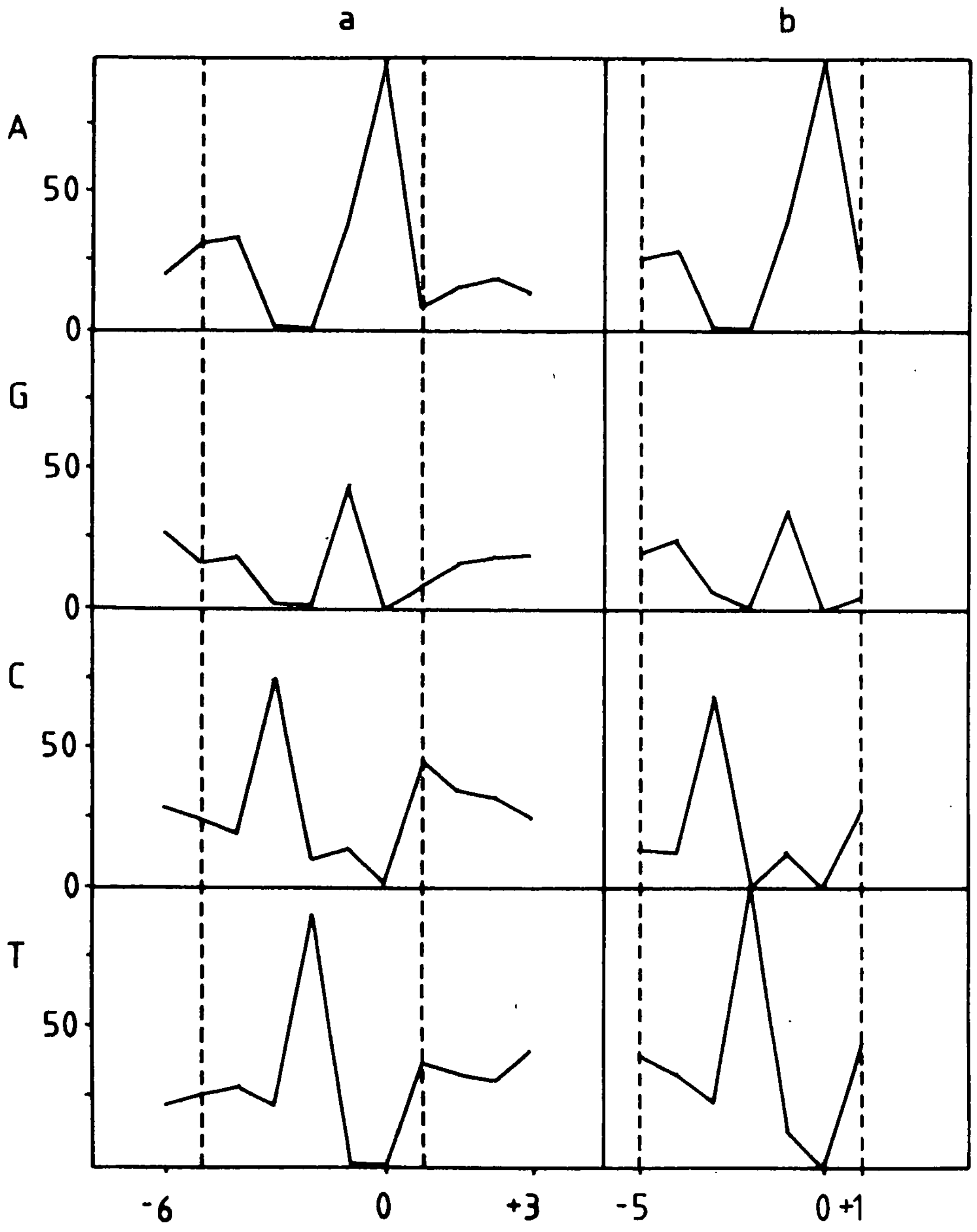

a
b
A
G
C
T
50
0
50
0
50
0
50
-6
0
+3
-5
0 +1

TABLE 11 Compilation of sequences around the intron 5′ splice site in pea seed storage protein genes.

		-3	-2	-1		+1	+2	+3	+4	+5	+6
LegA	IVS-1	A	A	G	:	G	T	T	A	C	T
	-2	A	G	G	:	G	T	G	A	G	C
	-3	A	A	A	:	G	T	A	T	G	T
LegB	IVS-1	A	A	G	:	G	T	T	A	C	T
	-2	A	G	G	:	G	T	G	A	G	C
	-3	A	A	A	:	G	T	A	T	G	T
LegC	IVS-1	A	A	G	:	G	T	T	A	C	T
	-2	A	G	G	:	G	T	G	A	G	C
	-3	A	A	A	:	G	T	A	T	G	T
LegD	IVS-1	A	A	G	:	G	T	T	C	G	T
	-2	A	G	G	:	G	T	G	A	G	A
	-3	A	A	A	:	G	T	A	C	C	A
LegJ	IVS-1	A	G	A	:	G	T	A	A	G	T
	-2	A	G	G	:	G	T	A	A	C	T
LegK	IVS-1	A	G	A	:	G	T	A	A	G	T
	-2	A	G	G	:	G	T	G	T	G	T
VicA	IVS-3	A	A	T	:	G	T	A	A	G	T
	-5	C	A	G	:	G	T	A	T	A	T
VicB	IVS-1	G	T	G	:	G	T	A	A	T	T
	-2	C	A	G	:	G	T	A	A	T	A
	-3	A	A	T	:	G	T	A	A	G	C
	-4	G	A	G	:	G	T	A	T	G	A
	-5	C	A	G	:	G	T	A	T	A	T
VicC	IVS-1	G	C	G	:	G	T	A	A	T	T
	-2	T	C	G	:	G	T	A	T	T	A
	-3	A	A	T	:	G	T	A	A	G	T
VicJ	IVS-1	G	T	G	:	G	T	A	A	T	G
	-2	C	A	G	:	G	T	A	A	T	A
	-3	A	A	T	:	G	T	A	A	G	C
	-4	G	A	G	:	G	T	A	T	A	A
	-5?	G	G	T	:	G	T	A	A	T	A
CvA	IVS-1	A	T	G	:	G	T	A	A	T	T
	-2	G	A	G	:	G	T	A	A	T	A
	-3	A	A	T	:	G	T	A	A	G	T
	-4	A	A	G	:	G	T	A	T	G	T
	-5	C	A	G	:	G	T	A	T	T	A
Totals	A	23	22	6	:	-	-	27	23	3	10
	G	7	9	24	:	36	-	5	-	18	1
	C	5	2	-	:	-	-	-	2	5	5
	T	1	3	6	:	-	36	4	11	10	20
Pea Consensus		A	A	G	:	G	T	A	A	G	T
Plant Consensus		C A	A	G	:	G	T	A	A	G	T

TABLE 12 Compilation of sequences around the intron 3′ splice site in pea seed storage proteins genes.

		-15					-10					-5				-1		+1
LegA	IVS-1	C	T	A	T	A	C	C	A	A	T	T	A	C	A	G	:	G
	-2	A	T	C	T	A	T	G	T	T	T	G	A	C	A	G	:	A
	-3	A	C	A	A	T	C	T	T	C	A	T	A	C	A	G	:	A
LegB	IVS-1	C	T	A	T	A	C	C	A	A	T	T	A	C	A	G	:	G
	-2	A	T	C	T	A	T	G	T	T	T	G	A	C	A	G	:	A
	-3	A	C	A	A	T	C	T	T	C	A	T	A	C	A	G	:	A
LegC	IVS-1	C	T	A	T	A	C	C	A	A	T	T	A	C	A	G	:	G
	-2	A	T	C	T	A	T	G	T	T	T	G	A	C	A	G	:	A
	-3	A	C	A	A	T	C	T	T	C	A	T	A	C	A	G	:	A
LegD	IVS-1	T	A	C	A	T	C	A	A	T	T	A	C	T	A	G	:	G
	-2									-								
	-3	A	C	A	A	T	T	T	T	C	A	T	A	C	A	G	:	A
LegJ	IVS-1	A	A	T	A	T	G	T	G	T	A	T	G	C	A	G	:	G
	-2	T	G	T	A	T	G	T	A	T	A	T	G	C	A	G	:	A
LegK	IVS-1	A	A	T	A	T	G	T	G	T	A	T	G	C	A	G	:	G
	-2	A	A	T	A	T	G	T	A	T	A	T	G	C	A	G	:	A
VicA	IVS-2	T	G	T	C	C	T	T	T	T	C	A	T	C	A	G	:	T
	-3	T	G	T	A	A	A	A	A	A	A	C	A	C	A	G	:	A
	-4	T	T	C	A	A	A	T	T	A	A	T	A	T	A	G	:	G
	-5	A	T	T	G	A	A	A	A	T	T	T	G	A	A	G	:	G
VicB	IVS-1	C	A	T	T	G	T	A	A	T	T	T	C	C	A	G	:	G
	-2	G	T	C	C	T	T	T	T	T	C	A	A	C	A	G	:	T
	-3	T	G	T	A	A	A	A	A	A	A	T	A	T	A	G	:	A
	-4	C	A	A	A	T	T	A	A	A	T	A	T	T	A	G	:	G
	-5	G	T	C	A	G	A	T	A	C	A	C	C	G	A	G	:	T
VicC	IVS-2	T	T	G	T	C	A	T	T	T	C	A	A	C	A	G	:	T
	-3	T	T	G	T	A	A	A	A	T	A	T	A	C	A	G	:	A
	-4	C	A	A	A	T	T	T	T	A	T	C	T	T	A	G	:	G
VicJ	IVS-1	A	A	T	T	G	C	A	A	T	A	T	G	C	A	G	:	G
	-2	T	G	T	A	A	T	T	T	T	T	A	A	T	A	G	:	T
	-3	G	T	T	C	A	A	A	A	T	A	T	A	A	A	G	:	A
	-4	C	A	A	A	T	T	T	A	A	T	T	T	T	A	G	:	G
CvA	IVS-1	T	A	A	T	A	T	G	T	A	C	T	A	C	A	G	:	G
	-2	A	T	C	C	T	T	C	T	T	C	T	A	C	A	G	:	G
	-3	T	T	T	T	C	G	C	A	G	A	T	A	T	A	G	:	A
	-4	T	G	C	T	T	A	A	A	T	T	T	T	T	A	G	:	G
	-5	A	T	T	G	A	A	A	A	T	T	T	G	A	A	G	:	G
Totals																		
	A	14	10	11	17	15	10	11	19	10	16	6	21	3	36	-	:	14
	G	3	6	2	3	3	5	4	2	1	-	3	7	1	-	36	:	17
	C	7	4	9	4	3	8	5	-	5	5	3	3	24	-	-	:	-
	T	12	16	14	12	15	13	16	15	20	15	24	5	8	-	-	:	5
Pea Consensus																		
		A	T	A	A	A	A	A	A	A	A	T	A	C	A	G	:	X
		T		T	T	T	T	T	T	T	T							
Plant Consensus																		
		T	T	T	T	T	T	T	T	T	T	T	G	C	A	G	:	G
					X			X	X	X	X							

TABLE 13 Compilation of sequences around the intron branch point in pea seed storage protein genes.

		-5					0	+1	SP
LegA	IVS-1	T	A	C	T	A	A	T	27
	-2	C	A	G	T	A	A	C	30
	-3	A	G	C	T	A	A	C	22
LegB	IVS-1	T	A	C	T	A	A	T	27
	-2	C	A	G	T	A	A	C	30
	-3	T	G	C	T	A	A	C	22
LegC	IVS-1	T	A	C	T	A	A	T	27
	-2	C	A	G	T	A	A	C	30
	-3	A	G	C	T	A	A	C	22
LegD	IVS-1	T	A	T	T	T	A	C	26
	-2					-			
	-3	G	A	C	T	T	A	A	28
LegJ	IVS-1	T	A	C	T	A	A	A	30
	-2	A	A	C	T	A	A	T	29
LegK	IVS-1	T	T	C	T	A	A	T	20
	-2	A	G	C	T	A	A	T	29
VicA	IVS-2	T	T	C	T	T	A	T	22
	-3	T	G	T	T	A	A	T	40
	-5	T	G	C	T	T	A	T	27
VicB	IVS-1	A	T	C	T	A	A	T	19
	-2	T	C	T	T	T	A	T	24
	-3	T	G	T	T	A	A	T	40
	-4	A	T	T	T	A	A	T	25
	-5	T	G	C	T	T	A	T	27
VicC	IVS-2	T	A	T	T	G	A	T	31
	-3	G	A	T	T	T	A	C	41
	-4	T	A	T	T	T	A	T	38
VicJ	IVS-1	A	T	C	T	T	A	A	15
	-2	C	T	C	T	T	A	C	28
	-3	C	T	T	T	T	A	T	22
	-4	A	G	C	T	A	A	A	25
CvA	IVS-1	G	T	T	T	A	A	T	13
	-2	G	A	T	T	G	A	A	29
	-3	G	T	T	T	A	A	T	44
	-4	T	G	T	T	A	A	T	24
	-5	T	T	C	T	A	A	T	23
Totals									
	A	8	14	-	-	22	35	5	
	G	5	10	3	-	2	-	-	
	C	5	1	19	-	-	-	9	
	T	17	10	13	35	11	-	21	
Consensus		T	X	Y	T	A	A	T	
						T			
Plant Consensus		T	T	Y	T	X	A	Y	
		X	X						

SP = Distance to 3′ splice site.

Overall the similarity shown between plant and animal genes in the 5' splice site and branch point regions, along with a lower degree of similarity at the 3' splice site, suggests possibly that the U-sn RNAs involved at the first two regions may show more sequence conservation than those involved at the latter.

A survey of pea seed storage protein genes was then compiled for these 3 regions in order to assess the degree of conformation to the corresponding plant consensus sequences:

i) 5' splice site - Table 11 shows a compilation for pea storage genes in this region along with the resulting implied consensus at each position, compared along with the plant consensus. The only apparent difference is at position -3, where the pea genes show a stronger preference for A, as opposed to an approx equal representation of A and C at this position in the plant sequence.

ii) 3' splice site - A similar compilation of data for this region is given in Table 12. Here good agreement is shown between the pea and plant consensuses from positions -15 to -5. However a major difference is seen to occur at -4, where instead of showing a prevalence of G's, the pea genes have A's in 21 out of the possible 36 places. One can say for sure whether this indicates a tolerance for A's as well as G's in this position in plant gene introns, or whether it in fact implies that the U-Sn RNA involved at the 3' splice site in peas is different to that occurring in other plants. However, the observation that broadbeans and pea U2-Sn RNAs (thought to be involved at the 3' splice site, see above) show near perfect homology throughout 121 nt at their 3' ends (Kiss et al., 1987) implies the latter theory may be incorrect. Another slight difference also appears to occur at position +1, with pea genes apparently tolerating A or G here, in contrast to a preference for only G here in the overall plant consensus sequence.

iii) The Branch point - Table 13 shows a compilation of sequences, in pea gene introns, of maximum homology to the consensus branch point sequence common to both animal and plant genes, i.e. C T Pu A Py (see above). The results show perfect agreement at positions -5, -4, -2, 0 and +1. Some disagreement can be observed at -3, where a general preference for Pyrimidines is shown in peas, rather than the more usual C in the plant consensus, and at -1, where the pea genes show a higher than expected proportion of T's, with correspondingly less G's. Such slight differences would appear fairly trivial in the light of evidence from *in vivo* and *in vitro* studies on a point mutation on the invariant A at position 0 in the yeast branch point sequence. Results showed surprisingly that mutation of this A to a C gave a high efficiency of branch formation (Vijayraghavan et al., 1986). This suggests that if a mutation of such an important position can be tolerated, then mutations elsewhere in the consensus might also be to some degree acceptable.

As a reference to pea genes in general, it is worth noting that the normal mechanisim of intron splicing via lariat formation at the predicted branch point, and cleavage at predicted splice junctions has actually been observed *in vivo* for *Leg* J intron + exon flanking sequences from pea, cloned behind an SP6 promotor (Brown et al., 1986)

Finally a point possibly worth considering is the efficiency with which the pre-mRNAs from foreign genes might be processed in a heterologous gene system. Should the mechanism prove to have some characterisitics unique to certain plant species, it could prove important in determining levels of foreign gene expression in transgenic plants.

4.1.7 Amino Acid and nucleic acid sequence alignments

Alignments were made, firstly, by eye, of all available amino acid sequence data for both 11S and 7S globulin legume seed storage proteins (the 11S alignment is shown in fig. 34 along with an example from a Brassica species - Cruciferin (*Bn Cr*) from *Brassica napus* (Simon et al., 1985) and that for 7S, partly taken from Doyle et al., 1986, in fig.35. Regions of conservation of amino acid residues are boxed, with various other features also indicated (see figure legends for details). Secondly, using the NUCALN computer programme, 3 nucleic acid sequence alignments for 11S genes were made from genomic and cDNA sequence data encoding a glycinim A_2B_{1a} subunit from soybean (GMA2B1a, Marco et al 1984) against sequence data from the legumin cDNA clones pDUB1 and pDUB3, encoding the basic subunit of a 60,000 Mr legumin polypeptide (Croy et al., 1982), and *Leg* A against *Leg* J, plus one for 7S genes, where *Vic* B was aligned against *Pvu* β . The results of these comparisons are illustrated graphically (see section 2.13.1) in fig. 36 (GMA2B1a vs Legumin cDNAs), fig. 37 (*Leg* A vs *Leg* J) and fig.38 (*Vic* B vs *Pvu* β).

The findings from these analyses were as follows:

4.1.7.1 11 S Globulins

The results of amino acid alignment (fig. 34) of the 11S seed storage proteins showed that many gaps, some of considerable size, had to be introduced to maximise homology in all exons. This homology can be seen to vary greatly between individual exons.

i) The signal peptide region - Here a good degree of conservation can be seen to exist between each protein, and also that each has an Ala residue next to the cleavage site - this appears to be in accordance with the findings of von Heijne (1985) who observed a proportionately high number of Ala residues in this position in a large number of proteins from both prokaryotes and eukaryotes.

ii) The rest of exon 1 - firstly, there appears to be no conservation at the residue occurring at the NH_2 termini of the mature polypeptides. However, throughout the rest of this exon, considerable sequence conservation can be seen to exist, with even *Bn Cr* , Cruciferin protein from a plant in a different extant subclass to the legumes, namely Dilleniidae as opposed to the Rosidae (Cronquist, 1981) showing some areas of homology to the legume proteins. This conservation appears greatest at residues from positions 46 to 73, where differences (see figure legend for the criteria by which difference is judged) occur at only 5 out of these 28 positions. Relatively few gaps were needed to achieve maximum alignment in this region.

iii) Exon 2 - referred to as such, but since some of the genes are known to lack the first intron present in *Leg* A, B, C, D (eg *Leg* J and *VfLe* B), this is merely a convenient term of reference.

This region also shows a large degree of conservation and again only a small number of gaps had to be inserted to maximise the alignment. *BnCr* shows some similarities again, but is distinct in its possession of a large insertion of sequence in the middle of the region (marked on the figure by an exclamation mark). Assuming that the present day genes coding for these proteins may have arisen from a common ancestral gene (purely on the basis of the amino acid homology which still appears to exist), this insertion must obviously have occurred subsequent to divergence from the ancestral gene.

iv) Exon 3 - This region corresponds to the C-terminus of the acidic polypeptide. Here the conservation of sequence present in the first two exons is seen to break down a short way into this exon consistent with observations made in section 4.1.2.1. Some degree of homology resumes midway from positions 288 to 307, but in general few residues are shared between the proteins, and large gaps have been inserted to achieve the degree of alignment shown. A full analysis and discussion on the evolutionary relationship between these proteins is presented later (see section 4.2.1), but purely by looking at the positioning of these gaps it would seem possible that *Leg* J and K, *VfLe* B and GM A3 B4, GMA5A4B3 and GMA1ABx are more closely related to each other then are *Leg* A, B, C, D and possibly GMA2B1a. Indeed gaps were needed in *Leg* J, K and *Vf Le* B to account for the 18 a.a. repeats present in *Leg* A B and C (starting at position 326). A gap spanning nearly all this region of repeats was needed in *Bn Cr*, and extended right to the cleavage point between the acidic and basic polypeptides. Here all the proteins show conservation starting at the Asn residue of the cleavage site and continuing uninterrupted for a further 8 residues, implying that structural integrity at this region, presumably on the surface of the protein molecule, is extremely important. Indeed, a good degree of conservation continues from here, corresponding to the N-terminal region of the basic polypeptide.

v) Exon 4 - This region can be seen to maintain the degree of conservation seen to commence at the N-terminal of the basic polypeptide in exon 3. The homology eventually tails off towards the C-terminal region, where a virtually in significant level of sequence conservation can be observed, implying that this region is of little importance to the overall structural integrity of each molecule.

Observing the alignment as a whole, other significant features emerge. Firstly, when the known positions of common introns are compared (bearing in mind the *Leg* J and *VfLe* B are lacking the first intron), they are seen to exactly correspond. However, little conservation of a.a sequence is observed around the intron sites - the greatest degree of conservation of this nature is shown around IVS-1, position 110, which is only so far known to occur in *Leg* A, B, C, and D.

Secondly, the observation that sizes of conserved blocks of a.a residues are often quite small suggests that evolutionary conservation is operating on units much smaller than whole exons, implying little relationship between exons and evolutionary domains.

Finally, identification of these evolutionarily important domains, partially permitted by the results of this alignment, along with potential 'hot spots' where amino acid substitution, insertion or deletion is apparently permitted, suggests that future protein engineering work might take advantage of these regions of the protein molecules that permit change without deleterious effect to their overall function.

Fig 34 Comparison of 11S storage protein amino acid sequences from three legumes plus one Brassica species; Data was predicted from either genomic or cDNA nucleotide sequence data and alignment was performed by hand. PS *Leg* J, K, A, B, C and D are legumin genes from *Pisum Sativum* (pea), VfLe B is a B type legumin gene from *Vicia faba* (broadbean), GMA3B4, GMA5A4B3, GMA1aBx and GMA2B1a are glycinin genes from *Glycine max* (soybean), and BNCr is Cruciferin from *Brassica napus* (see text for references). The sequences have been aligned with gaps (represented by dashes) included where required to give maximum homology. A large insertion in BNCr is not shown, but its position is indicated by an exclamation mark. Intron positions are indicated by arrows. The N-terminal of each mature acidic polypeptide is indicated by a colon, as is the border between each acidic and basic polypeptide. An asterisk denotes the C terminal of each mature basic polypeptide. Amino acid residues shared by at least one storage protein in each of the three legume genera are boxed. Additionally, due to observations made on the radically differing nature of the *Leg* A (so called A type legumins) and *Leg* J (so called B type legumins) type proteins (see Section 4.1.2), it was decided, for the purposes of this analysis only, to treat *Leg* J and *Leg* K genes as a further genera, to be boxed according to the above mentioned criteria. BNCr was included in boxes were amino acid conservation was observed Conservative subsitutions (V=I=L; E=D; Y=F; S=T) were considered as shared residues.

```
                                                                                                                          V
PSLegJ    MIRVTMS-KPF--LSLLSLSLLLF-AS-ACLA:--TSSEFDRLNQCQLDSINALEPDHRVVSEAGLTETWNPNHPELKCAGVSLIRRTIDPNGLHLPSFSPSPQLIFIIQGKGVLGLSFP
VFLeB          MS-KPF--LSLLSLSLLLF-TS-TCLA:--TSSEFDRLNQCRLDNINALEPDHRVESEAGLTETWNPNHPELRCAGVSLIRRTIDPNGLHLPSYSPSPQLIYIIQGKGVIGLTLP
PSLegA         MA-KLL-ALS-LSFCFLLL-GG-CF-A:--LREQ-PQQNECQLERLDALEPDNRIESEGGLIETWNPNNKQFRCAGVALSRATLQRNALRRPYYSNAPQEIFIQQGNGYFGMVFP
PSLegB         MA-KLL-ALS-LSFCFLLL-GG-CF-A:--LTEQ-PQQNECQLERLDALEPDNRIESEGGLIETWNPNNKQFRCAGVALSRATLQRNALRRPYYSNAPQEIFIQQGNGYFGMVFP
PSLegC         MA-KLL-ALS-LSFCFLLL-GG-CF-A:--LTEQ-PQQNECQLERLDALEPDNRIESEGGLIETWNPNNKQFRCAGVALSRATLQRNALRRPYYSNAPQEIFIQQGNGYFGMVFP
PSLegD         MATKLL-ALS-LSLCFLLF-SS-CF-A:--LREQ-S*NNECQLERLNALGPDNCIESEG*LVETWNPNNRQFRCVGVTLSRTTLQPNAFRRPYYSNAPQKIYIQQGNGYFGIVFP
GMA3B4         MG-KPFFTLSLSSLCLLLL-SSACF-A:-ITSSKF---NECQLNNLNALEPDHRVESEGGLIETWNSQHPELQCAGVTVSKRTLNRNGSHLPSYLPYPQMIIVVQGKGAIGFAFP
GMA5A4B3       MG-KPF-TLSLSSLCLLLL-SSACF-A:-ISSSKL---NECQLNNLNALEPDHRVESEGGLIQTWNSQHPELKCAGVTVSKLTLNRNGLHSPSYLPYPRMIIIAQGKGALGVAIP
GMA1aBx        MA-KLV----FSLCFLLFSGC-CF-A:FSSREQ-PQQNECQIQKLNALKPGNRIESEGGLIETWNPNNKPFQCAGVALSRCTLNRNALRRPSYTNGPQEIYIQQGKGIFGMIYP
BNCr                                        QQFP--NECQLDQLNALEPSHVLKAEAGRIEVWDHHAPQLRCSGVSFVRYIIESKGLYLPSFFSTAKLSSVAKGEGLSGRVVL

                                            40                                      80                                       120
                                                                                    V
PSLegJ    GCPETYEEPRSSQSRQG--SR-QQQGDSHQKVRRFRKGDIIAIPSGIPYWTYNHGDEPLVAISLLDTSNIANQLDSTPRVFYLGGNPETEFPETQEEQQGRH-----------RQKHSYPV
PSLegK                                        GIPYWTYNHGDEPLVAISLLDTSNIANQLDSTPRVFYLGGNPETEFPETQEEQQGRH-----------RQKHSYPV
VFLeB     GCPQTYQEPRSSQSRQG--SR-QQQPDSHQKIRRFRKGDIIAIPSGIPYWTYNNGDEPLVAISLLDTSNIANQLDSTPRVFYLGGNPEVEFPETQEEQQERH-----------QQKHSLPV
PSLegA    GCPETFEEPQESEQGEG--R-RY--RDRHQKVNRFREGDIIAVPTGIVFWMYNDQDTPVIAVSLTDIRSSNNQLDQMPRRFYLAGNHEQEFLQYQHQQGGKQ-----------EQENEGNN
PSLegB    GCPETFEEPQESEQGEG--R-RY--RDRHQKVNRFREGDIIAVPTGIVFWMYNDQDTPVSAVSLTDIRSTNNQLDQMPRRFYLAGNHEQEFLRYQHQQGGKQ-----------EQENEGNN
PSLegC    GCPETFEEPQESEQGEG--R-RY--RDRHQKVNRFREGDIIAVPTGIVFWMYNDQDTPVIAVSLTDIRSTNNQLDQMPRRFYLAGNHEQEFLRYQHQQGGKQ-----------EQENEGNN
PSLegD    GCPETFEELQESEQSEG--R-RY--RDSHQKVNRFREGDIIAVPSGVVFWMYNDQDTPVIAISLTETGSSNNQLDQMPRRFYLAGNAL--------------------------------
GMA3B4    GCPETFEKPQQQSSRRGSRSQ-QQLQDSHQKIRHFNEGDVLVIPLGVPYWTYNTGDEPVVAISPLDTSNFNNQLDQNPRVFYLAGNPDIEHPETMQQQQQQKSHGGRKQGQHRQQEEE---
GMA5A4B3  GCPETFEEPQEQSNRRGSRSQKQQLQDSHQKIRHFNEGDVLVIPPSVPYWTYNTGDEPVVAISLLDTSNFNNQLDQTPRVFYLAGNPDIEYPETMQQQQQQKSHGGRKQGQH-QQEEEE-
GMA1aBx   GCSSTFEEPQQPQQRGQ--SS--RPQDRHQKIYNFREGDLIAVPTGVAWWMYNNEDTPVVAVSIIDTNSLENQLDQMPRRFYLAGNQEQEFLKYQQEQGGHQSQKGKHQQ--EEENE---
BNCr      -CAETFQDSSV-FQPSG---G--SPQDNHQKVEHIRTGDTIATHPGVAQWFYNDGNQPLVIVSVLDLASHQNQLDRNPRPFYLAGNNPQGQVWIEGREQQ------PQH----------N
                                 !
                                           160                                      200     .                                240

PSLegJ    GRRSGHHQQEEESEQQNEGNSVLSGFSSEFLAQTFNTE--EDTAKRLRSPRDERSQIVRVEGGLRIII--KGR--N-EEEKEQSHSHS-------------------------
PSLegK    GRRSGHHQQEEESEEQNEGNSVLSGVSSEFLAQTFNTE--EDTAKRLRSPRDERSQIVRVEGGLRIINPKGK--EEEEEKEQSHSHS-------------------------
VFLeB     GRRGGQHQQEEESEEQKDGNSVLSGFSSEFLAQTFNTE--EDTAKRLRSPRDKRNQIVRVEGGLRIINP---------------------------------------------
PSLegA    IFSGFKRDFLEDAFNV-------------------NRH--IVDRLQGRNEDEEKGAIVKVKGGLSII--SPP--E-KQARHQRGSRQE-EDEDEE------------K-QPRHQRGSRQE
PSLegB    IFSGFKRDFLEDAFNV-------------------NRH--IVDRLQGRNEVEEKGAIVKVKGGLSVI--SPP--E-KQARHQRGSRQE-EDEDEE------------K-QPRHQRGSRQE
PSLegC    IFSGFKRDFLEDAFNV-------------------NRH--IVDRLQGRNEDEEKGAIVKVKGGLSII--SPP--E-KQARHQRGSRQE-EDEDEE------------K-QPRHQRGSRQE
PSLegD    -----------------------------------KKH--IVDKLQGRNKDEEKGAIVKVKGGLSII--SPP--E-RQPHHQRGSRQ---------------------------------
GMA3B4    ------------------GGSVLSGFSKHFLAQSFNTN--EDTAEKLRSPDDERKQIVTVEGGLSVI--S-P--K-V---------QEQEDEDEDEDEEYG--RTPSY-PPRRPSHGKHE
GMA5A4B3  ------------------GGSVLSGFSKHFLAQSFNTN--EDIAEKLESPDDERKQIVTVEGGLSVI--S-P--K-V---------QEQQDEDEDEDEDDEDEQTPSH-PPRRPSHGKRF
GMA1aBx   ------------------GGSILSGFTLEFLEHAFSVDKQIAKNLQGENEGEDKGAIVTVKGGLSVI--KPPTDE-QQ---QR-P-QE-EEDDDE------------EEQP--QCVGKRF
GMA2B1a                                                                                E-EDDDDE------------EEQP--QCV----
BNCr      ILNGFTPEVLAKAFKI-------------------DVR--TAQQLQNQQ--DNRGNIIRVQGPFSVI--RPP--L-RS---QR-P-QE-EV---------------------------

                                           280                                      320                                      360
```

```
PSLegJ    --------------------------------------------------HREEKEEEEEEEEEDEEEKQRS-EER-----KN:GLEETICSAKIRENIADAARADLYNPRAGRISTANS
PSLegK    ---------------------------------------------------HREEEEEEE---EDEE-KQRS-EER-----KN:GLEETICSAKIRENIADAAGADLYNPRAGRISTANS
VFLeB     -------------------------------------------------------EGQQEEEEEEEEEKQRS-EER-----KN:GLEETICSAKIRENIAQPARADLYNPRAGSISTANS
PSLegA    EEEDEDEER-QPRH---------------------------------Q-RRRGEEEEEDKKERGGSQKGKSR-RQG-----DN:GLEETVCTAKLRLNIGPSSSPDIYNPEAGRIKTVTS
PSLegB    EEEDEDEER-QPRH---------------------------------Q-RRRGEEEEEDKKERRGSQKGKSR-RQG-----DN:GLEETVCTAKLRLNIGPSSSPDIYNPEAGRIKTVTS
PSLegD    ---DEDEER-QPCH---------------------------------Q-RRRGEEEEEDEKERHCSQKGKSR-VHG-----DN:GLEETICTAKLRQNIGSSSSPNIYNPVAGRIKTVTS
PSLegC    EEEDEDEER-QPRH---------------------------------Q-RRRGEEEEEDKKERGGSQKGKSR-RQG-----DN:GLEETVCTAKLRLNIGPSSSPDIYNPEAGRIKTVTS
GMA3B4    DDEDEDEEEDQPRPDHPP-----------------------------Q-RPSRPEQQE--------PRGRGC--QT-----EN:GVEENICTMKLHENIARPSRADFYNPKAGRISTLNS
GMA3A4B3  QDEDEDEDEDKPRPSRPSQGKRNKTGQDEDEDEDEDQPRKSREWRSKKTQPRRPRQEE--------PRERGC--ET-----EN:GVEENICTLKLHENIARPSRADFYNPKAGRISTLNS
GMA1aBx   --------------------------------------------------------GKDK----HCQ----RPRGSQSKSRRN:GIDETICTMRLRHNIGQTSSPDIYNPQAGSVTTATS
GMA2B1a   --------------------------------------------------------ETDK----GCQ---------QSKS-RN:GIDETICTMRLRQNIGQNSSPDIYNPQAGSITTATS
BNCr      ----------------------------------------------------------------------------------N:GLEETICSARCTENLDDPSNADVYKPQLGYISTLNS

                                                                400                                      440                  480
                    V
PSLegJ    LTLPVVRYLRLSAEYVRLYRNDIYAPHWNINANSLLYVIRGEGRVRI-RSCELPTWTMFDNKL-RKGHLVVVPQNFVVSEQAGEEEGLEYVVFKTNDRAAVSHVQQVFRATPSEVLANAF
PSLegK    LTLPVLRYLRLSAEYVRLYRNDIYAPHWNINANSLLYVIRGEGRVRI---VNFQGDAVFDNKL-RKGQLVVVPQNFVVSEQAGEEEGLEYVVFKTNDRAAVSHVQQVLRATPAEVLANAF
VFLeB     LTLPILRYLRLSAEYVRLYRNGIYAPHWNINANSLLYVIRGEGRVRI---VNSQGNAVFDNKV-RKGQLVVVPQNFVVAEQAGEEEGLEYLVFKTNDRAAVSHVQQVFRATPADVLANAF
PSLegA    LDLPVLRWLKLSAEHGSLHKNAMFVPHYNLNANSIIY--ALKGRARL-QVVNCNGNTVFDGKL-EAGRALTVPQNYAVAAKSLSDRFS-YVAFKTNDRAGIARLAGTSSVINNLPLDVVA
PSLegB    LDLPVLRWLKLSAEHGSLHKNAMFVPHYNLNANSIIY--ALKGRARL-QVVNCNGNTVFDGKL-EAGRALTMPQNYAVAAKSLSDRFS-YVAFKTNDRAGIARLAGTSSVINNLPLDVVA
PSLegC    LDLPVLRWLKLSAEHGSLHKNAMFVPHYNLNANSIIY--ALKGRARL-QVLNCNGNTVFDGKL-EAGRALTVPQNYAVAAKSLSDRFS-YVAFKTNDRAGIARLAGTSSVINNLPLDVVA
PSLegD    LDLPLFRWLKLIAEHGSLHKNAMVFPHYNLNANCVIY--TLKGRARL-QIVNCNGNTVSDG-------------------------------------KAAIARLAGTSSTLNAMPVDVIA
GMA3B4    LTLPALRQFGLSAQYVVLYRNGIYSPDWNLNANSVTM-TRGKGRVR---VVNCQGNAVFDGEL-RRGQLLVVPQNPAVAEQGGEQGLE-YVVFKTHHNAVSSYIKDVFRVIPSEVLSHSY
GMA5A4B3  LTLPALRQFQLSAQYVVLYKNGIYSPHWNLNANSVIYVTRGQGKVR---VVNCQGNAVFDGEI-RRGQLLVVPQNFVVAEQAGEQGFE-YIVFKTHHNAVTSYLKDVFRAIPSEVLAHSY
GMA1aBx   LDFPALSWLRLSAEFGSLRKNAMFVPHYNLNANSIIY---LNGRA-LIQVVNCNGERVFDGELQE-GRVLIVPQNFVVAARSQSDNFE-YVSFKYNDTPMIGTLAGAKSLLNALPEEVIQ
GMA2B1a   LDFPALWLLKLSAQYGSLRKNAMVFPHYTLNANSIIY--ALNGRALV-QVVNCNGERVFDGELQE-GGVLIVPQNPAVAAKSQSDNFE-YVSFKTNDRPSIGNLAGANSLLNALPEEVIQ
BNCr      YDLPILRFLRLSALRGSIRQNAMVLPQWNANANAVLY--VTDGEAHV-QVVN--GDRVFDGQV-SQGQLLSIPQGFSVVKRATSEQFR-WIEFKTNANAQINTLAGRTSVLRGLPLEVISN

                                                      520                                      560                      600

PSLegJ    GLRQEQVTELKLSGNRGPMVHPR-SQSQSH*----------
PSLegK    GLRQEQVTELKLSGNRGPLVHPQ-SQSQSH*----------
VFLeB     GLRQEQVTELKLSGNRGPLVHPQQSQSQSN*----------
PSLegA    ATFNLQRNEARQLKSNNPFKFLVPARESENRASA*------
PSLegB    ATFNLQRNEARQLKSNNPFKFLVPARQSENRASA*------
PSLegC    ATFNLQRNEARQLKSNNPFKFLVPARQSENRASA*------
PSLegD    ATFNLQRSEARQVKSNNPFKFLVPPRESENKASA*------
GMA3B4    NGGQSQV---RQLKYQGNSGPLLVNP*--------------
GMA5A4B3  NLRQSQV---SELKYEGNWGPLVNPESQQGSPRVKVA*---
GMA1aBx   HTFNLKSQQARQIKNNNPFKFLVPPQESQKRAVA*------
GMA2B1a   HTFNLKSQQARQVKNNNPFSFLVPPQESQRRAVA*------
BNCr      GYQIS-LEEARRVKFNTIETTLTHSSGPASYGGPRAKADA*

                                                      640
```

In comparison with the a.a. alignment, the profiles of nucleic acid homology for *Leg* A vs *Leg* J (fig.37) and 60,000 Mr Legumin cDNA vs GMA2B1a (fig.36) confirm many of the findings listed above, especially that conservation of sequence is greatest in the N-terminal half of the acidic polypeptide, that there is a highly variable region in the C terminal region of this polypeptide, and that there is again good homology shown at the N-terminal of the basic polypeptide. The *Leg* A vs *Leg* J profile also seems to imply that sequence conservation at regions bordering the introns common to the two genes is strong - this tends to refute the theory of exon shuffling (see section 4.1.2.1), where one would expect homology to be strongest in the central region of the exon and to break down at intron border regions, thus allowing functional units to be transferred in the 'shuffling' process without damage occurring to them.

4.1.7.2. <u>7S Globulins</u>

Overall, fewer gaps were needed for this alignment (see fig. 35) than were used in that for the 11S globulins. Again homology varied between individual exons.

i) <u>The signal peptide</u> - Some degree of conservation can be observed here, but the cleavage site and residues around it do not appear to show uniformity. Unlike the 11S globulin proteins, none of these signal peptides conform to the described pattern of residues at this site observed in a range of eukaryotic and prokaryotic proteins (von Heijne, 1983). There also appears to be a broad similarity in the lengths of each signal peptide, with only the convicilin (CVA) leader showing a difference in this region. The difference between CVA and Psa3, an a.a. sequence predicted from pea convicilin cDNA (Casey et al., 1984) in this region is of interest. Since the copy number for convicilin genes in the haploid pea genome has been estimated as 1 for pea (Domoney and Casey, 1985), exact correspondence of a.a. sequences between CVA and Psa 3 might have been expected. The difference observed may be due either to damage of the Psa3 cDNA at its 5' end, giving an altered predicted a.a. sequence, or to the fact that the cDNA encoding Psa3 and the CVA genomic clones were isolated from different pea lines, Birte and Dark Skinned Perfection respectively. Another possible reason is that there may in fact be more than the estimated 1 gene copy of convicilin in haploid pea genome. (Recent findings now in fact support the last theory in confirming that there are indeed 2 convicilin genes in the pea genome, Bown et al., 1988).

ii) The remaining section of exon 1 - As is the case with the 11S proteins a high degree of conservation can be observed in this region. Indeed from position 47 to 116, 42 of these 69 residues are conserved, with blocks of conservation occurring over as many as 8 and 9 residues. The N-terminal region shows no conservation however, and gaps are needed in order to achieve the alignment for the rest of this region. Indeed, two of the proteins contain large insertions of amino acid sequence not shown in this figure. *Gma* α ' has a 174 a.a insert 7 residues from the N-terminus of the mature protein, and CVA has one of 121 a.a's 4 residues away from its N-terminus. Studies have shown (Bown et al., 1988) that no homology exists between these two inserts, implying that each is the result of a separate insertion event during gene evolution. It has been observed that the insert present in *Gma* α ' is also found in a gene encoding the α subunit of β -conglycinin (Schuler et al., manuscript in preparation), and since it is not present in any of the other genes represented here, represents an insertion event that occurred subsequent to the divergence of β conglycinin α and α ' subunit genes from that encoding *Gma* β (Doyle et al., 1986) - see section 4.2.2. for further discussion on this point.

iii) Exon 2 - Here the conservation is seen to be much less pronounced, and large gaps were needed to take account of additional sequence in *Gma* α ' and *Gma* β - this section suggest possibly that insertions in equivalent regions of the other proteins might be tolerated in future protein engineering work.

iv) Exon 3 - Extensive conservation can be seen in this short exon, and only two small gaps had to be inserted in the alignment procedure.

v) Exon 4 - Conservation in this region breaks down after the first 15 residues, and after this point, many gaps had to be inserted to identify the only other area of homology existing in the centre of this exon.

vi) Exon 5 - Interestingly, conservation appears to exist only at the border regions of this exon. The central region required the insertion of gaps in many of the sequences and would appear to present another potential target for protein engineering.

vii) Exon 6 - Unlike the C-terminal region of the basic polypeptides in the 11S globulins, this region shows a surprisingly high degree of conservation extending right to the C-terminal residue of many of the polypeptides, implying that this region is of some importance to the proper functioning of each protein.

Considering the alignment across all regions, other features can be observed. Firstly, identification of proteolytic cleavage sites in the pea vicilins (Lycett et al., 1983) enabled the identification of one at position 375 in both 47,000 Mr Polypeptides Psal and *Vic* J, which can be seen to be a region of low conservation between the 7S polypeptides. This might imply that individual regions of conservation unique to the 7S polypeptides of one species but not to others, may still be subjected to strong evolutionary constraint and thus although apparently in a highly variable region with respect to polypeptide of other species, would not be good site for modification in the species where they are conserved.

Secondly, a number of potential N-glycosylation sites have been identified in these polypeptides (Doyle et al., 1986). Of the 3 indentified in *Gma* α ', only 1 is common to any of the other polypeptides - it is at position 418-420 and occurs also in *Gma* α Gma β and Psa 1, but is not present however in any of the other polypep^ptides. A site identified at position 424-426 in both *Pvu* α and *Pvu* β is not present in any of the other polypeptides. If these sites identified are functional, these findings seem to imply that positioning of these sites is not crucial to polypeptide function, and also possibly that the original purported ancestral polypeptide may not have been glycosylated at all (or conversely it had several sites which have been lost to varying degrees during the evolutionary process which produced the modern day descendant polypeptides.

Finally, looking at this data in conjunction with the nucleic acid sequence alignment profile, (fig. 38) of *Vic* B vs *Pvu* β it is clear that the position of each intron is conserved across the whole range of 7S genes. However, since areas of a.a. residue conservation are often seen to span intron positions (e.g. introns 1, 4 and 5), and also that from the nucleic acid sequence alignment profile, areas of sequence homology continue from one exon into the next, a similar situation to that suggested for the 11S proteins seems to exist (see section 4.1.7.1.). Here it was suggested that such findings refuted the idea of exon shuffling, and this again appears to be the case for 7S proteins. One could not conclude from this data that the exons represent the discreet functional domains required for the shuffling theory.

Fig. 35 Comparison of 7S storage protein amino acid sequences from 3 legumes. Data has been predicted from either genomic or cDNA nucleotide sequence data. The alignment was compiled by addition to and slight modification of, that published by Doyle et al., 1986, and was performed by hand. Pvu a and Pvu b are alpha and beta type (respectively) phaseolin genes from *Phaseolus vulgaris*, Gma a', Gma a, and Gma b are alpha^1, alpha and beta subunit genes respectively for β -conglycinin of *Glycine max*, Psa 1 and Psa 2 represent 47 and 50,000 Mr vicilin cDNAs respectively, from *Pisum sativum* (Lycett et al., 1983), Psa 3 is a cDNA clone representing a convicilin gene from *Pisum Sativum* (Casey et al., 1984), PSCVA is a genomic clone representing a pea convicilin gene and PS Vic J, A, B, C are all vicilin genomic clones from pea (references not documented here can all be found in the text). The sequences have been aligned with gaps (represented by dashes) included where required to give maximum homology. The large insertion in exon 1 of Gma a' is not shown. The N-terminal of each mature polypeptide is indicated by a colon, and the C-terminus by the arrows. Amino acid residues shared by at least one storage protein in each of the three genera are boxed; conservative substitutions (V=I=L; E=D; Y=F; S=T) are considered shared residues.

```
                                                                                                                      V
Pvua     M---MRARVPLLLLLGILFLASLSAS---F-:ATSLREEEESQD----NPFYFNSDNSWNTLFKNQYGHIRVLQRFDQQSKRLQNLEDYRLVEFRSKSETLLLPQQADAELLLVVRSGSAI
Pvub     M---MRARVPLLLLLGILFLASLSAS---F-:ATSLREEEESQD----NPFYFNSDNSWNTLFKNQYGHIRVLQRFDQQSKRLQNLEDYRLVEFRSKPETLLLPQQADAELLLVVRSGSAI
Gmaa'    M---MRARFPLLLLLGVVFLASVSVS---F-:GIAYWEKREPRRHKNKNPFHFNSKRFQ-TLFKNQYGHVRVLQRFNKRSQQLQNLRDYRILEFNSKPNTLLLPHHADADYLIVILNGTAI
Gmab     M---MRVRFPLLVLLGTVFLASVCV---S-:LKVREDEN--------NPFYFRSSNSFQTLFENQNVRIRLLQRFNKRSQQLENLRDYRIVQFQSKPNTILLPHHADADFLLFVLSGRAI
Psa1     M---LLAIA---------FLASVCV---SS:RSDQE-----------NPFLFKSNRFQ-TLYENENGHIRLLQKFDKRSKIFENLQNYRLLEYKSKPHTLFLPQYTDADFILVVLSGKAT
Psa3                         ER---SS:ES-QE-----RR---NPFLFKSNKFL-TLFENENGHIRLLQRFDKRSDLFENLQNYRLVEYRAKPHTIFLPQHIDADLILVVLSGKAL
PSCvA    MATTVKSRFPLLLFLGIIFLASVCVTYASS:ES-QE-----RR----NPFLFKSNKFL-TLFENENGHIRRLQRFDKRSDLFENLQNYRLVEYRAKPHTIFLPQHIDADLILVVLNGKAI
PSVicJ   M---AATPIIKLMLLAIAFLASVCV---SS:RSDQE-----------NPFIFKSNRFQ-TLYENENGHIRLLQKFDKRSKIFENLQNYRLLEYKSKPHTLFLPQYTDADFILVVLSGKAT
PSVicB   M---KAS-FPLLMLMGISFLASVCV---SS:RSDPQ-----------NPFIFKSNKFQ-TLFENENGHITLLQKFDQRSKIFENLQNYRLLEYKSKPHTIFLPQHTDADYILVVLSGKAI
PSVicC   M---KAS-FPLLMVLGIAFLASVCV---SS:RSDQD-----------NPFIFESKRFQ-TLFENEKVHIRLLQKFDQRSKIFENLQNYRLLEYKSKPHTIFLPQQTDADFILVVLS

                                                 40                                        80                                        120
                                                                                   V                               V

Pvua     LVLVKPDDRREYFFLTQSDNPIFS-------------DHQKIPAGTIFYLVNPDPKEDLRIIQLAMPVNKP-QIH-DFFLSSTEAQQSYLQEFSKHILEASFNSKFEEINRVLFE-EEGQ
Pvub     LVLVKPDDRREYFFLT-SDNPIFS-------------DHQKIPAGTIFYLVNPDPKEDLRIIQLAMPVNNP-QIH-EFFLSSTEAQQAYLQEFSKHILEASFNSKFEEINRVLFE-EEGQ
Gmaa'    LTLVNNDDRDSYNLQS-GDALRVPAGTTFYVVNPDNDENLRMIAGTTFYVVNPDNDENLRMITLAIPVNKP-EAFESFFLSSTQAQQSYLQGFSKNILEASYDTKFEEINKVLFGREEGQ
Gmaa                                                                                                        SRNNILEASTDTKFEEINKVLFSREEGQ
Gmab     LTLVNNDDRDSYNLHP-GDAQRIPYGTTYYLVNPHD-----------------HDNLKIINLAIPVNKP-GRYDDFFLSSTQAQQSYLQGFSHNILETSFHSEFEEINRVLFGEEEEQ
Psa1     LTVLKSNDRNSFNLER-GDAIKLP------------------AGSIAYFANRDDNEEPRVLDLAIPVNKPGQLQ-SFLLSGTQNQKSSLSGFSKNILEAAFNTNYEEIEKVLLEQQEQE
Psa2                                                                                                                  DNAEIEKILLEENEKF
Psa3     LTVLSPNDRNSYNLER-GDTIKLP------------------AGTISYLVNQDDEEDLRLVDLVIPVNGPGKPE-AFDLAKNKNQYLR--GFSKNILEASYNTRYETIEKVLLEEQEK-
PSCvA    LTVLSPNDRNSYNLER-GDTIKIP------------------AGTTSYLVNQDDEEDLRVVDFVIPVNRPGKPE-AFGLSENKNQSLR--GFSKNILEASLNTKYETIEKVLLEEQEKK
PSVicJ   LTVLKSNDRNSFNLER-GDTIKLP------------------AGTIAYLANRDDNEDLRVLDLAIPVNKPGQLQ-SFLLSGTQNQPSLLSGFSKNILEAAFNTNYEEIEKVLLEQQEQE
PSVicA                                                                            SFLLSGNQNQQSYLSGFSKNILEASFNTDYEEIEKVLLEEHEKE
PSVicB   LTVLNPDDRNSENLER-GDTIKLP------------------AGTIAYLVNRDDNEELRVLDLAIAVNRPGQLQ-SFLLSGNQNQQNYLSGFSKNILEASFNTDYEEIEKVLLEEHEKE
PSVicC                                                                            SFLLSGNQNQQSILSGFSKNILEASFNTDYEEIEKILLEEHEKE

                                                 160                                       200                                       240

                                                                                                           V
Pvua     QEEGQQ----------EG-VI--VDSEQIEELSKHAKSSSRKSHSKQDNTIGNEFGNLTERTDNSLNVILSSIEM--------------------EGALFVPHYYSKAIVIL-VNEGEAH
Pvub     Q---------------EG-VI--VDSEQIKELSKHAKSSSRKSLSKQDNTIGNEFGNLDERTDNSLNVLISSIEM--------------------EGALFVPHYYSKAIVIL-VNEGEAH
Gmaa'    QQGEERLQ--------ES-VI--VSKKQIEELSKHAKSSSRKTISSEDKPFNL--GSRNPITSNKLGKLFEITQR-K--PQLRDLDVFLSVVDMNEGALFLPHFNSKAIVVL-INEGEAN
Gmaa     QQGEQRLQ--------ES-VI--VSKKQIRTLSKRAKSSSRKTISSEDKPFNL                          PQLRDLDIFLSIVDMNEGALLLPHFNSKAIVIL-INEGDAN
Gmab     EQQ-------------EG-VI--VSKKQIEQLSKRAKSSSRKTISSEDEPFNL--RSRNPIYSNNFGKFFEITPE-N--PQLRDLDIFLSSVDINEGALLLPHFNSKAIVIL-INEGDAN
Psa1     PQHRRSLKDRRQEINEEN-VI--VSKDQIEELSKNAKSSSKKSVSSESGPFNL--RSRNPIYSNKFGKFFEITPE-N--QQLQDLDIFVNSVDIKVGSLLLPHYNSRAIVIV-VTEGKGD
Psa2     THHAAGLRDKRQQSQEKN-VI--VSKKQIEELSKNAKSSSKKSVSSRSEPFNL--RSSDPITSNQYGKFFEITPK-N--PQLQDLDIFVNYVEIKEGSLLLPHYNSRAIVIV-VNEGKGD
Psa3     -----DRK-RRQQGEETDAIV-KVS
PSCvA    PQQLRDRK-RTQQGEERDAII-KVSREQIEELRKLAKSSSKKSLPSEFEPFNL--RSHKPEYSNKFGKLFEITPE-KKYPQLQDLDILVSCVEIN-GALMLPHYNSRAIVVLLVNEGKGN
PSVicJ   PQHRRSLKDRRQEINEEN-VIVKVSREQIEELSKNAKSSSKKSVSSESGRFNL--RSRNPIYSNKFGKFFEITPEKN--QQLQDLDIFVNSVDIKEGSLLLPHYNSRAIVIVTVTEGKGD
PSVicA   TQHRRSLKDK                                                                                    GSILLPHYNSRAIVIVTVNEGKGD
PSVicB   TQHRRSLKDKRQQSQEEN-VIVKLSRGQIEELSKNAKSTSKKGVSSRSEPFNL--RSRGPIYSNEFGKFFEITPEKN--PQLQDLDIFVNSVEIKEGSLLLPHYNSRAIVIVTVNEGKGD
PSVicC   THHRRGLRDKR*QSQEKN-VIVKVSNKQIEELSKNA                                                          GSILLPHYNSRAIVIVTVNEGKGG

                                                 280                                       320                                       360
```

```
                                                                                         V
Pvua     VELVGPKGNKE---T---------------LEFESYRAELSKDDVFVIPAAYPVAIKATSNVNFTGFGINANNNNRNLLAGKTDNVISSIGALDGKDVLGLTFSGDGDEVMKLINKQSGS
Pvub     VELVGPKGNKE---T---------------LEFESYRAELSKDDVFVIPAAYPVAIKATSNVNFTGFGINANNNNRNLLAGKTDNVISSIGALDGKDVLGLTFSGDGDEVMKLINKQSGS
Gmaa'    IELVGIKEQQQ---R--------QQ-QEEQPLEVRKYRAELSEQDIFVIPAGYPVVVNATSDLNFFAFGINAENNQRNFLAGSKDNVISQIPQ-----VQELAFPHSAKDIENLIKSQSES
Gmaa     IELVGLKEQQQ---E--------QQ-QEEQPLEVRKYRAELSEQDIFVIPAGYPVVVNATSNLNFFAIGINAENNQRNFLAGSQDNVISQIPQ-----VQELAFLGSAQAVEKLLKNQRES
Gmab     IEVVGIKEQQQ---K--------QK-QEEEPLEVQKYRAELSEDDVFVIPAAYPFVVNATSNLNFLAFGINAENNQRNFLAGEKDNVVRQIEQ-----VQELAFPGSAQDVERLLKKQRES
Psa1     FELVGQRNENQ---GKENDKEEEQ-EEETSKQVQLYRAKLSPGDVFVIPAGHPVAINASSDLNLIGLGINAENNERNFLAGEEDNVISQVEP-----VKELAFPGSSHEVDR
Psa2     FELVGQRNENQQGLREEDDEEEEQREEETKNQVQSYKAKLTPGDVFVIPAGHPVAVRASSNLNLLGFGINAENNQRNFLAGEEDNVISQIQQ-----VKDLTFPGSAQEVDRLLENQKQS
PSCvA    LELLGLKNEQQ----------EREDRKERNNEVQRYEAKLSPGDVVIIPAGHPVAISASSNLNLLGFGINAKNNQRNFLSGSDDNVISQIEN----PVKELTFPGSSQEVNELIKNQKQS
PSVicJ   FELVGQRNENE---GKENDKEEEQ-EEETSKQVQLTRAKLSPG
PSVicA   FELVGQRNENQQEQRKEVDEEEEQGEEEINKQVQNYKAKLSSGDVFVIPAGHPVAVKASSNLDLLGFGINAENNQRNFLAGDEDNVISQIQR----PVKELAFPGSAQEVDRILENQKQS
PSVicB   FELVGQRNENQQEQRKEDDEEEEQGEEEINKQVQNYKAKLSSGDVFIPAGHPVAVKATSNLDLLGFGINAENNQRNFLAGDEDNVISQIHR
PSVicC   FELVGQRNENQQGLREEDDEEEEQREEETKNQVQSYKAKLTPGDVFVI                                                       DLTFPGSAQEVDRLLENQKQS

                                       400                                  440                                     480
```

```
Pvua     YFVDAHHHQ-QEQQKGSHQQEQQKGRKG-------AFVY*
Pvub     YFVDAHHHQ-QEQQKGR-----------KG-------AFVY*
Gmaa'    YFVDA-----QPQQK----EEGNKGRKGPLSSILRAF-Y*
Gmaa     YFVDA-----QPKKK----EEGNKGRKGPLSSILRAF-Y*
Gmab     YFVDA-----QPQQK----EEGSKGRKGPPPSILGAL-Y*
Psa2     YFANA-----QPQQR----ETRSQEIKEHLYSILGAF*
PSCvA    HFASA-----EPEQK----EEESQRKRSPLSSVLDSF-Y*
PSVicA   HFADA-----QPQQR----EEGSRETED
PSVicC   YFANA-----QPQQR----ETRSQEIKEHLYSILGAF*

                            500                520
```

A comparison of the trends shown in both the 11S and 7S amino acid sequences alignments reveals certain characteristics consistent with those described in a comparison of structural similarities between legumin and vicilin storage proteins from legumes (Argos et al., 1985). The main common feature between 11S and 7S proteins is the variability shown by both in the central region of the polypeptides, also the site of large insertions in some of the legumin proteins and some smaller insertions in an number of vicilin proteins. Additionally, the residues in both these types of insertion are mainly Asp and Glu.

Similar high degrees of conservation were also observed in this study for the NH_2 and COOH terminal regions of both 11S and 7S polypeptides.

Fig. 36 Percentage homology plot for most of the basic subunit from a pea legumin 60,000 Mr cDNA against soybean glycinin gene GMA2B1a (alignment from Momma et al., 1985) according to the method outlined in section 2.13.1. The dotted line indicates the significant level of homology. The arrow indicates the end of the coding sequence.

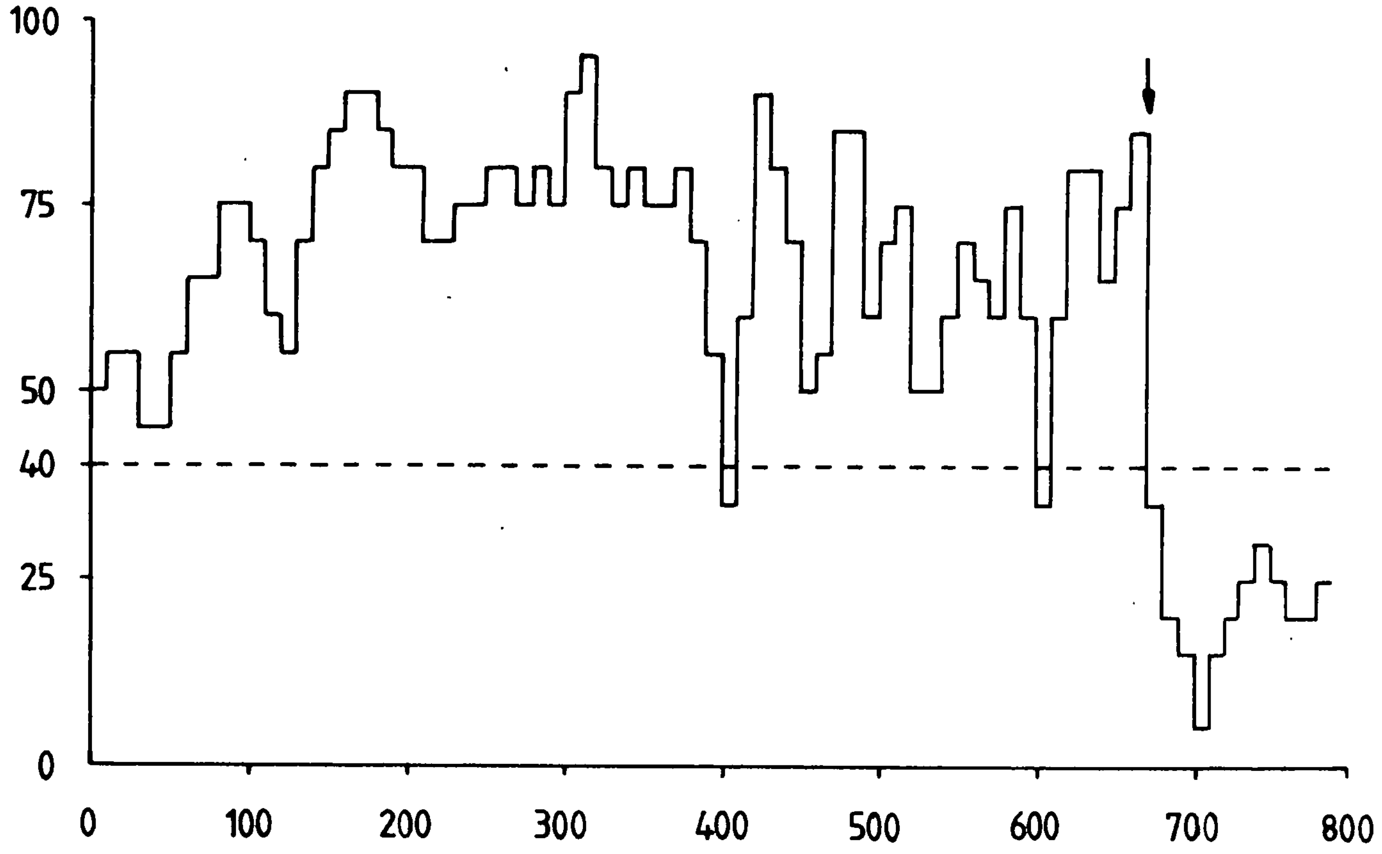
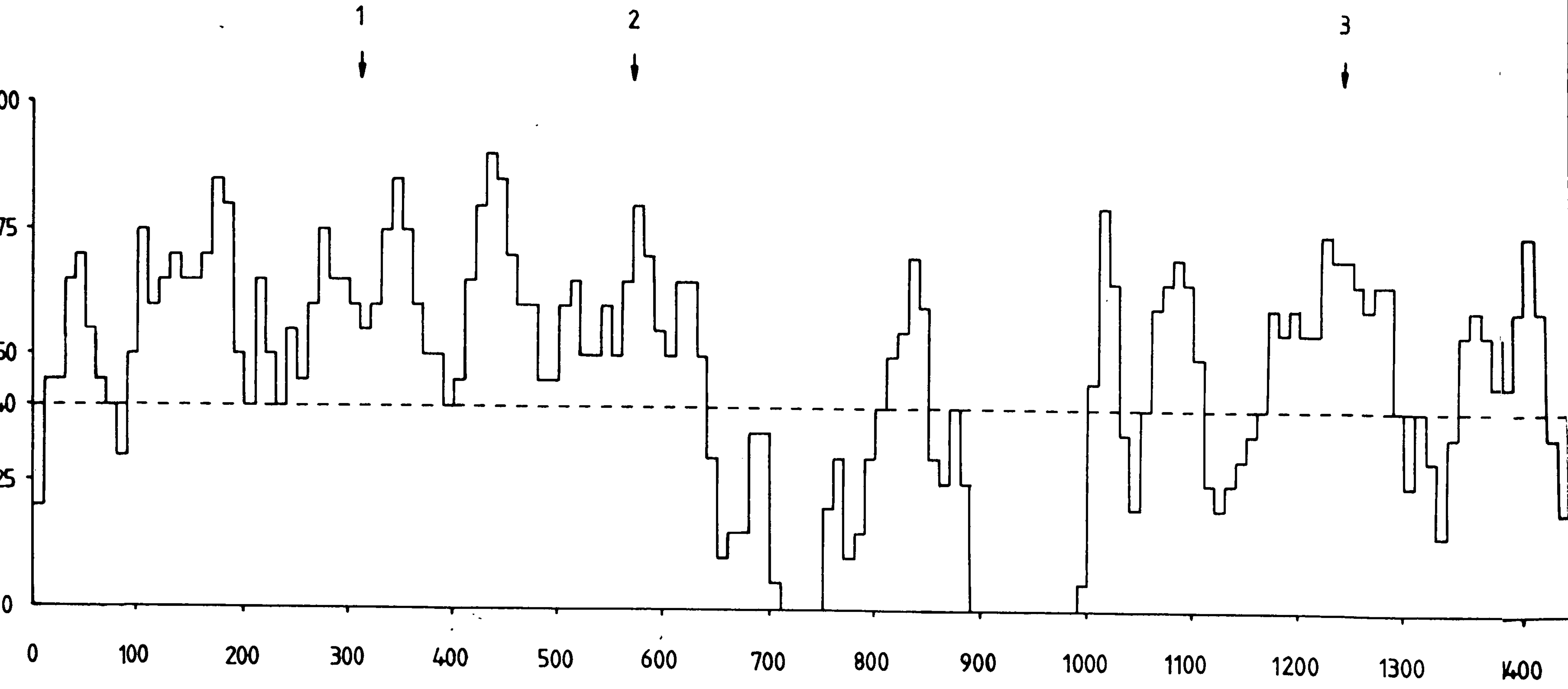
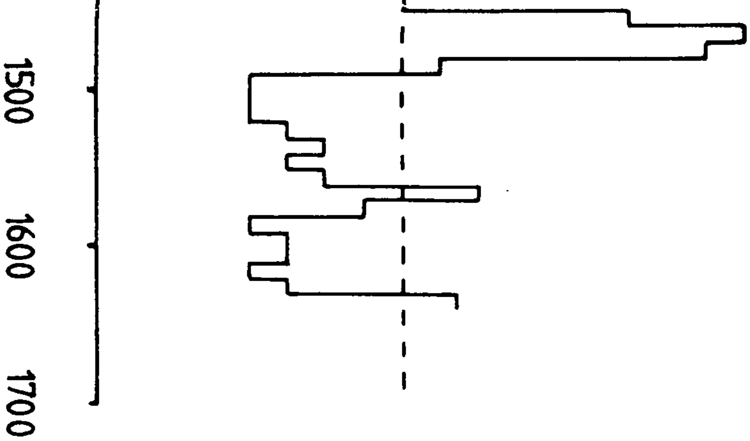

100
75
50
40
25
0
0
100
200
300
400
500
600
700
800

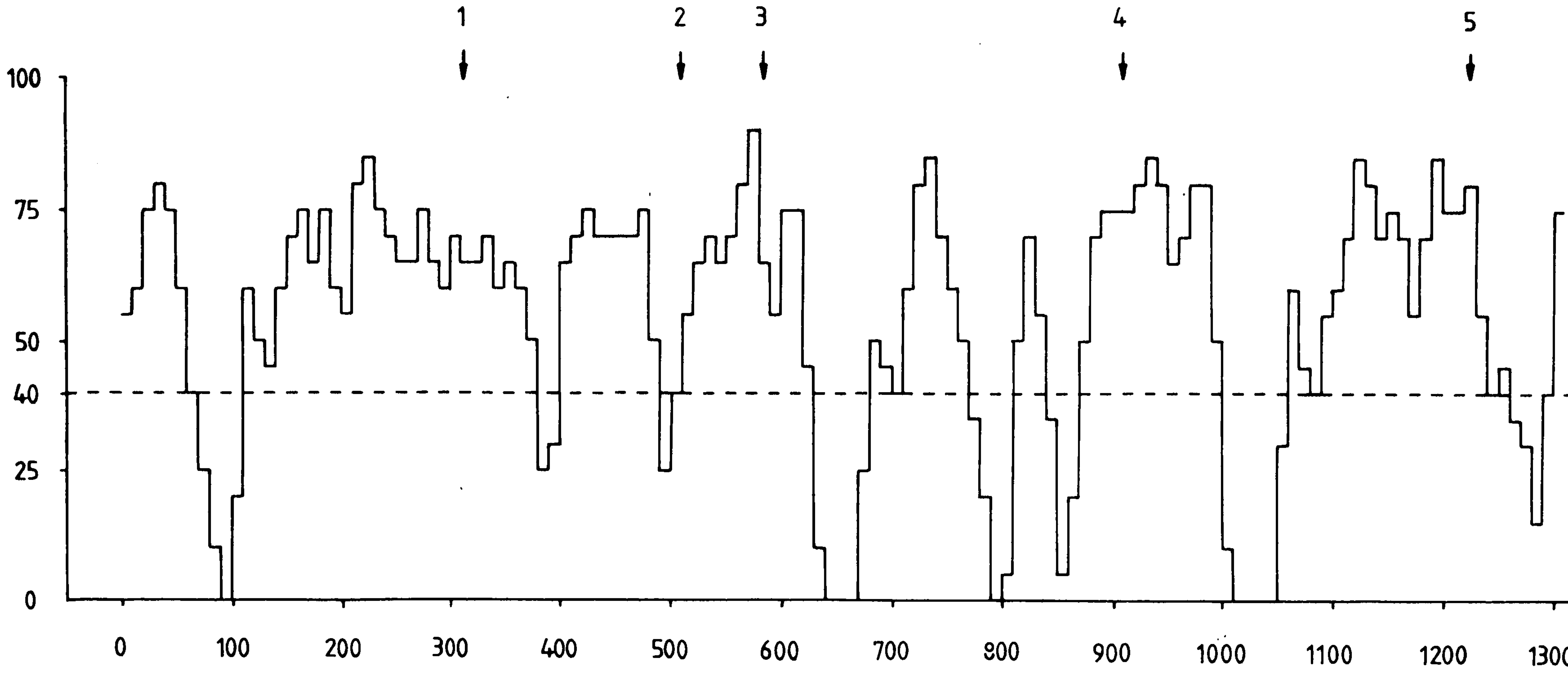
Fig. 37 Percentage homology plot for *Leg* A vs *Leg* J coding sequences, according to the method of section 2.13.1. The dotted line indicates the significant level of homology, and numbered arrows indicate positions of introns 1, 2, and 3 in *Leg* A, and of introns 1 and 2 in *Leg* J.


1
2
3
00
75
50
40
25
0
0
100
200
300
400
500
600
700
800
900
1000
1100
1200
1300
1400



1500
1600
1700

Fig. 38 Percentage homology plot for pea vicilin gene *Vic* B against Phaseolin gene *Pvu* β from *Phaseolus vulgaris* (see text for reference), according to the method outlined in section 2.13.1. The dotted line indicates the significant level of homology, and numbered arrows indicate positions of the 5 corresponding introns of the two genes. (coding sequences only compared)


1
2
3
4
5
100
75
50
40
25
0
0
100
200
300
400
500
600
700
800
900
1000
1100
1200
1300

4.2 Analysis of Evolutionary relationships amongst the 11S and 7S seed storage proteins

In order to determine possible patterns by which the genes coding for 11S and 7S storage proteins might have evolved, dendrograms for both classes were derived using the method outlined in section 2.13.3. Fig. 39 shows the two dissimilarity matrices obtained, and then using the single link classification algorithm on each of these matrices, the resulting dendrograms for 11S and 7S storage proteins are shown in Fig. 40. These dendrograms effectively represent orthologous comparisons to give phenetic relationships between the polypeptides concerned (see section 1.1.4).

Fig. 39 DC matrices for Single Link cluster analyses on both 11S and amino acid alignments (for method, see section 2.13.3). Numbering follows:

No.	A = 11S genes	B = 7S genes
1	*Leg* J (Legumin, pea)	Pvu a (phaseolin, type a, *P.vulgaris*)
2	*Leg* K (Legumin, pea)	Pvu b (phaseolin, type b, *P.vulgaris*)
3	*VfLe* B(Legumin, *Vicia faba*)	Gma a (β conglycinin, a' subunit, soybea
4	*Leg* A (Legumin, pea)	Gma a (β conglycinin, a subunit, soybean
5	*Leg* B (Legumin, pea)	Gma b (β conglycinin, b subunit, soybean
6	*Leg* C (Legumin, pea)	Psa 1 (vicilin 47000 Mr, pea)
7	*Leg* D (Legumin, pea)	Psa 2 (vicilin 50,000 Mr, pea)
8	GMA3B4 (Glycinin, soybean)	Psa 3 (convicilin, pea)
9	GMA5A4B3 (Glycinin, soybean)	CVA (convicilin, pea)
10	GMA1aBx (Glycinin, soybean)	*Vic* J (vicilin 47,000 Mr pea)
11	GMA2B1a (Glycinin, soybean)	*Vic* A (vicilin 50,000 Mr pea)
12	BNCr (Cruciferin, *Brassica napus*)	*Vic* B (vicilin 50,000 Mr pea)
13	-	*Vic* C (vicilin 50,000 Mr pea)

- for references, see text.

A

	1	2	3	4	5	6	7	8	9	10	11	12
1	0.000											
2	0.060	0.000										
3	0.095	0.090	0.000									
4	0.617	0.619	0.665	0.000								
5	0.620	0.622	0.667	0.013	0.000							
6	0.620	0.622	0.667	0.006	0.012	0.000						
7	0.590	0.548	0.569	0.139	0.154	0.149	0.000					
8	0.378	0.384	0.373	0.515	0.520	0.517	0.535	0.000				
9	0.397	0.389	0.367	0.545	0.547	0.547	0.558	0.142	0.000			
10	0.626	0.618	0.570	0.321	0.328	0.328	0.360	0.540	0.471	0.000		
11	0.697	0.652	0.661	0.419	0.431	0.431	0.381	0.604	0.478	0.133	0.000	
12	0.653	0.661	0.621	0.596	0.593	0.598	0.594	0.584	0.621	0.587	0.582	0.000

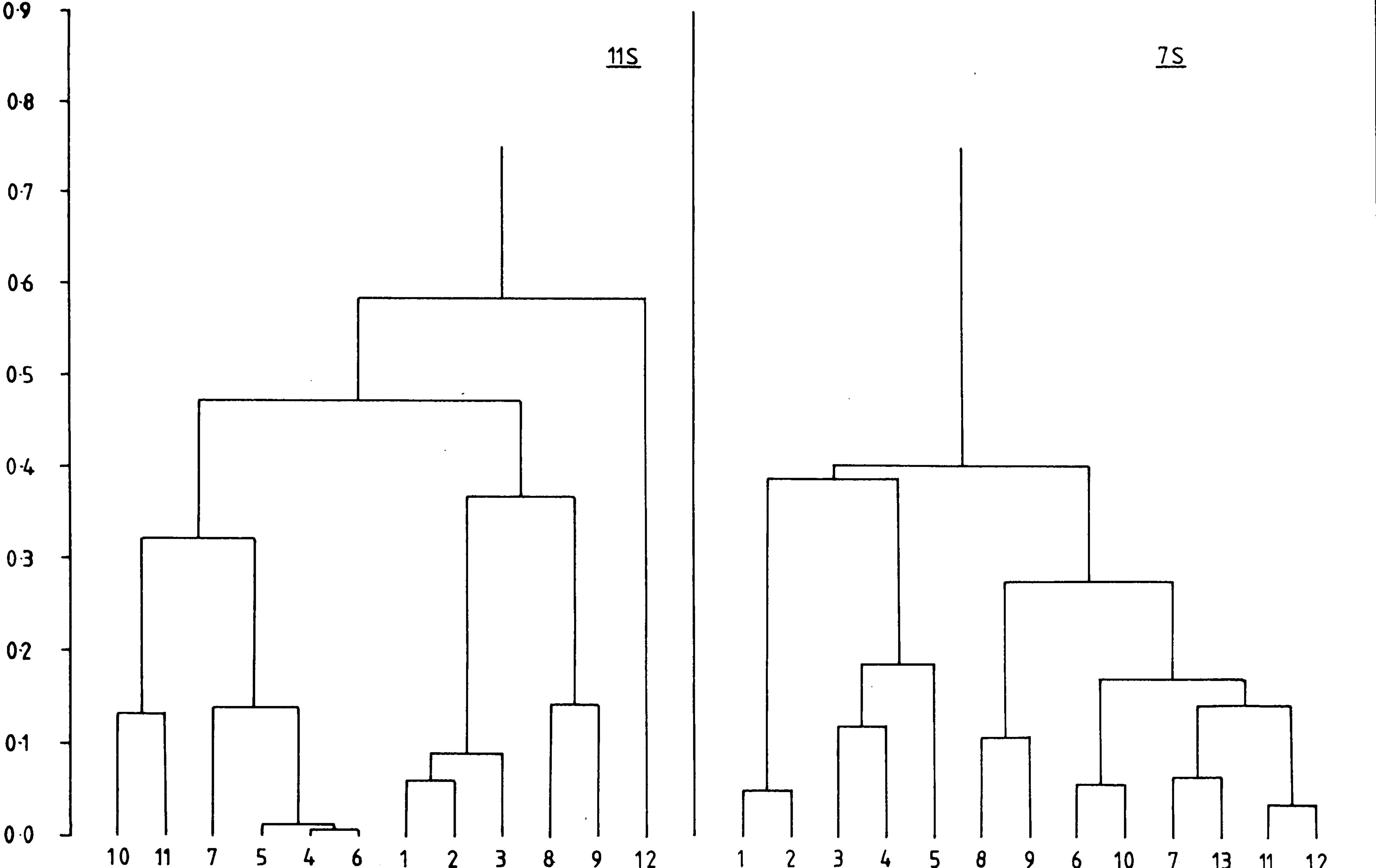
B

	1	2	3	4	5	6	7	8	9	10	11	12	13
1	0.000												
2	0.050	0.000											
3	0.413	0.413	0.000										
4	0.402	0.387	0.120	0.000									
5	0.457	0.437	0.265	0.186	0.000								
6	0.540	0.530	0.444	0.426	0.406	0.000							
7	0.539	0.533	0.406	0.448	0.402	0.252	0.000						
8	0.545	0.530	0.508	0.543	0.500	0.308	0.441	0.000					
9	0.546	0.525	0.452	0.450	0.441	0.310	0.391	0.107	0.000				
10	0.557	0.555	0.480	0.475	0.439	0.056	0.283	0.295	0.323	0.000			
11	0.503	0.514	0.431	0.453	0.423	0.190	0.213	0.341	0.350	0.214	0.000		
12	0.518	0.503	0.432	0.413	0.414	0.179	0.205	0.275	0.288	0.170	0.035	0.000	
13	0.508	0.496	0.476	0.467	0.435	0.196	0.065	0.331	0.368	0.219	0.182	0.142	0.000

Fig. 40 Dendrograms produced from the DC-matrices of fig.39 for 11S and storage proteins according to the method outlined in section 2.13 Numbering as follows:

No.	A = 11S	B = 7S
1	*Leg* J (Legumin, pea)	Pvu a (phaseolin, type a, *P.vulgaris*)
2	*Leg* K (Legumin, pea)	Pvu b (phaseolin, type b, *P.vulgaris*)
3	*VfLe* B (Legumin, *Vicia faba*)	Gma a' (β conglycinin, a' subunit, soybe
4	*Leg* A (Legumin, pea)	Gma a (β conglycinin, a subunit, soybean
5	*Leg* B (Legumin, pea)	Gma b (β conglycinin, b subunit, soybean
6	*Leg* C (Legumin, pea)	Psa 1 (vicilin 47,000 Mr, pea)
7	*Leg* D (Legumin, pea	Psa 2 (vicilin 50,000 Mr, pea)
8	GMA3B4 (Glycinin, soybean)	Psa 3 (convicilin, pea)
9	GMA5A4B3 (Glycinin, soybean)	CVA (convicilin, pea)
10	GMA1aBx (Glycinin, soybean)	*Vic* J (vicilin 47,000 Mr pea)
11	GMA2B1a (Glycinin, soybean)	*Vic* A (vicilin 50,000 Mr pea)
12	*BNCr* (Cruciferin, Brassica napus)	*Vic* B (vicilin 50,000 Mr pea)
13		*Vic* C (vicilin 50,000 Mr pea)

- for references, see text


0·9
0·8
0·7
0·6
0·5
0·4
0·3
0·2
0·1
0·0
11S
7S
10
11
7
5
4
6
1
2
3
8
9
12
1
2
3
4
5
8
9
6
10
7
13
11
12

Construction of dendrograms using data extracted from amino acid sequence does have its limitations (see section 1.1.4.). However, in order to perform the analysis at nucleic acid sequence level, individual pairwise sequence comparisons for 12 11S genes and 13 7S genes would have had to be carried out using the method described by Perler et al., 1980 to estimate the divergence time between each pair of genes. Such an extremely lengthy set of calculations was considered beyond the scope of this project. Also, this process could not have provided absolute times of divergence for each of the genes anyway, since no event of known date was available (such as the time of speciation for pea and soybean) to facilitate calibration of an evolutionary clock for these genes.

However, despite the acknowledged drawbacks of the method used to derive the dendrograms, they do still represent a systematic analysis of the data, and as such reveal some interesting features relating to how the present day 11S and 7S genes may have originated.

4.2.1 11S genes

Firstly, from this dendogram, an interesting pattern of gene divergence with respect to speciation emerges. Clearly the genes coding for *Leg* A, B and C show the least divergence. However, the branching pattern indicates that soybean GMA2B1a and GMA1aBx (plus not surprisingly pea *Leg* D) are more closley related to this class of gene then they are to the other glycinin genes from soybean, GMA3B4 and GMA5A4B3. Indeed these latter two soybean genes are shown to be more closely related to the group containing not only pea *Leg* J and K, but also *Vf Le* B from broadbean. Hence one can conclude that the classification of 11S genes into A and B classes (Baumlein et al., 1986) is correct. Class A represents *Leg* A, B, C, D plus GMA2B1a and GMA1aBx, and Class B *Leg* J and K, *Vf Le* B, plus GMA3B4 and GMA5A4B3. Since different species are represented in the classes, it is clear that speciation between soybean, broadbean and pea must have occurred subsequent to the duplication event which gave rise to the two different classes of gene. This is consistent with observations on the relatedness of *Leg* J and K with *Vf Le* B, GMA5A4B3 and GMA3 B4 and of *Leg* A, B, C, D, with GMA2B1a and GMA1aBx made in section 4.1.2.1. However, the suggestion made in section 4.1.7.1, that GMA1aBx might be more closely related to the *Leg* J, K class than to that representing *Leg* A etc., appears to have been erroneous.

Furthermore, on the basis of lengths of branches between the internodes, it appears firstly that *Vf Le* B, *Leg* J and *Leg* K are more closely related to each other than any is to GMA3B4, and that these soybean genes diverged from each other before the duplication(s) which gave rise to *Vf Le* B and *Leg* J and K. On the other hand within the A-type subfamily, the two soybean genes GMA2B1a and GMA1aBx are clearly more closely related to one another than to any of *Leg* A, B, C or D. The dendrogram also shows that the origin of legumin and glycinin subclasses occurred at roughly the same time in evolution. The distance shown by *Leg* D to the rest of the members of its subclass may be misleading due to its nature as a pseudogene (see section 4.1.1.) subjected to relatively little functional constraint compared to *Leg* A, B or C.

Not surprisingly, since it represents a gene from a plant in a different extant subclass to the rest of the genes analysed, *Bn Cr* shows only a distant relatedness, and appears to have originated by a very distant duplication event, with the other gene from this duplication going on to further diverge and duplicate to produce the 11 S genes present in legume species. This is consistent with the observation made in section 4.1.7.1. of an insert in exon 2 unique to *Bn Cr* also suggesting its early divergence, and overall agrees with the part of the classification of Angiosperms by Cronquist (1981) placing *Brassica napus* (rape) in a different extant subclass to that of soybean and pea.

4.2.2 7S genes

Here a different picture emerges to the one presented by the 11S genes. In this case, none of the many pea genes are more closely related to genes from other species than they are to themselves. Each one appears to have originated via duplication events which occurred *after* speciation from soybean and frenchbean. However it is interesting to note that soybean and frenchbean appear more closely related to one another than either does to pea - however spacing of the internodes, which indicates this fact, is fairly close, indicating that the speciation process which gave rise to Soybean and frenchbean occurred only shortly after speciation from pea. This finding is consistent firstly with the taxonomy suggested by Polhill (1981), and secondly with observations on levels of homology between the third exon of genes from each of these species (Doyle et al., 1986).

Several other observations made earlier in this discussion are consistent with the findings from the 7S dendrogram. Firstly, from sequence comparisons in section 4.1.3.5. that *Vic* C is most closely related to *Vic* B, (rather than to *Vic* J, CVA *Pvu* β or *Gma* α '), and that CVA had arisen from a duplication event which occurred prior to subsequent duplications giving rise to *Vic* C,B and J. These sequence comparisons also agree with the conclusion made above that soybean and frenchbean are more closely related to each other than either is to pea. Also, in relation to the soybean 7S genes, the suggestion made in section 4.1.7.2. on the basis of the placing of unique insertions in *Gma* α and *Gma* α ', that these two genes are more closely related to one another, than either is to *Gma* β is borne out by findings from the dendrogram. *Gma* α and *Gma* α ' appear to have arisen from a duplication event after divergence from *Gma* β . Similarly the suggestion that pea CVA was a product of the first duplication of a 7S protogene in peas also appears to be correct. A point of interest related to this is the surprisingly large evolutionary distance implied from the dendrogram for pea CVA and Psa 3, another convicilin sequence. This distance appears to confirm the suspicion raised earlier (section 4.1.7.2) that there may indeed be 2 gene copies for convicilin in the haploid pea genome.

Finally, data from the dendrogram also appears to concur with the conclusions of Casey et al., 1984, in that it implies that pea, soybean and french bean 7S proteins diverged prior to the divergence of vicilin and convicilin in pea.

On a general note, conclusions drawn from both 11S and 7S dendrograms agree with an earlier suggestion (made from comparisons of representatives from a wide range of Angiosperm 11 and 7S globulin seed storage proteins) that these two protein types are descendants of two ancestral genes (Borroto and Dure 1987).

4.3 Codon Usage and its potential significance in pea storage protein genes

4.3.1 Patterns of codon usage

It is now widely recognised that usage of synonymous codons in a given organism is non random, and this codon usage pattern has been shown to have characteristically different patterns between animal and bacterial genes (Grantham et al., 1981) and also between animal and plant genes (Lycett et al., 1983b). In *E.coli* codon usage has been shown to correlate with abundance of corresponding t RNAs for each synonymous codon (Ikemura 1985), and a similar relationship has been suggested for mammalian systems from observations on rabbit and human β globin mRNAs (Kafatos et al.,1977). This latter suggestion has been questioned however, since each β globin mRNA in fact has a different pattern of codon usage, making a link with levels of iso-accepting t RNAs seem unlikely (Jukes and Lester King, 1979).

Table 14 shows a comparison between codon usage patterns observed in animals and plants in general (Lycett et al., 1983b) alongside those observed for several individual pea seed storage protein genes. Several features of significance appear to emerge when the table is studied closely. Firstly, consistent with patterns observed in animal and plant genes in general, there appears to be a distinct lack of preference for codons containing the dinucleotide CG - e.g. Ser TCG,

TABLE 14

% Synonymous Use

	Animal	Plant	*Leg* A	*Leg* J	*Vic* B (50k)	*Vic* J (47k)	CVA
Phe TTT	31.7	44.7	36.3	53	57.2	55.6	69.6
TTC	68.3	55.3	63.6	47	42.8	44.4	30.4
Len TTA	1.9	9.1	0	13.0	16.3	21.9	22.4
TTG	8.9	24.6	16.6	17.3	25.6	26.8	22.4
CTT	8.9	27.8	33.3	28.2	39.5	34.1	24.4
CTC	26.7	16.0	21.4	23.9	7.0	9.7	8.1
CTA	6.9	14.4	19.0	8.6	9.3	4.9	10.2
CTG	46.5	8.0	9.5	8.6	2.3	2.4	12.2
Ile ATT	28.2	42.2	45.4	45.8	40.0	35	34.6
ATC	61.5	33.0	27.3	37.5	20.0	40	26.9
ATA	10.2	24.8	27.3	16.6	40.0	25	38.4
VAL GTT	13.2	33.9	36.0	41.3	33.3	15.8	37.9
GTC	30.8	13.2	8.0	10.3	4.1	21.0	10.4
GTA	7.3	19.8	20.0	10.3	37.5	36.9	27.5
GTG	48.5	33.0	36.0	37.9	25.0	26.3	24.2
Ser TCT	15.0	24.8	22.5	25	44.1	37.5	28.5
TCC	24.6	13.9	9.6	12.5	17.6	18.8	4.7
TCA	12.3	24.2	22.5	16.6	17.6	18.8	33.3
TCG	2.7	1.8	6.4	2.0	0	0	2.3
AGT	16.4	16.7	9.6	31.25	17.6	15.6	16.6
AGC	28.7	18.8	29.0	12.5	2.9	9.3	14.3
Pro CCT	30.4	30.3	29.1	40.7	40.0	20	34.7
CCC	36.9	24.2	20.8	7.4	6.7	13.3	26.9
CCA	21.7	37.3	33.3	44.4	53.3	60.0	30.7
CCG	10.8	8.1	16.6	7.4	0	6.7	7.7
Thr ACT	25.0	34.4	35.7	39.1	16.6	50.0	31.3
ACC	46.6	29.7	35.7	8.6	33.3	35.7	25.0
ACA	18.3	32.8	28.5	39.1	41.6	14.3	43.7
ACG	10.0	3.1	0	13.0	8.3	0	0
Tyr TAT	30.3	57.7	30.7	66.3	44.5	55.5	60
TAC	69.7	42.3	69.3	33.4	55.5	44.4	40
Ala GCT	32.5	39.1	45.9	32.1	41.2	56.2	33.4
GCC	44.1	20.5	13.5	25.0	29.4	25.0	28.5
GCA	16.3	37.1	37.8	32.1	29.4	18.8	33.3
GCG	7.0	3.3	2.7	10.7	0	0	4.8
His CAT	32.3	59.4	50.0	41.1	25.0	25.0	33.4
CAC	67.7	40.6	50.0	58.9	75.0	75.0	66.6

	Animal	Plant	*Leg* A	*Leg* J	*Vic* 3 (50k)	*Vic* J (47k)	CVA
Gln CAA	26.3	75.4	65.7	77.4	70.3	91.0	90
CAG	73.7	24.6	34.3	22.6	29.7	9.0	10
Asn AAT	22.2	47.0	42.R	30.7	58.8	51.7	59.4
AAC	77.8	53.0	57.8	69.2	41.2	48.3	40.6
Lys AAA	27.9	55.5	45.8	35.2	67.8	64.3	54.7
AAG	72.1	44.5	54.2	64.8	32.2	35.7	45.3
Asp GAT	40	60.5	54.2	31.3	81	62.5	63.6
GAC	60	39.5	45.8	68.8	19	37.5	36.4
Glu GAA	38.1	54.0	54	61.4	41.5	55.8	64.9
GAG	61.9	46.0	46	38.6	58.5	44.2	35.1
Cys TGT	43.4	44.4	57.1	66.6	100*	100*	50**
TGC	56.6	55.6	42.9	33.4	0	0	50
Arg CGT	17.4	15.6	12.5	22.2	10.0	21.0	18
CGC	23.9	5.2	14.5	11.1	10.0	5.2	8
CGA	8.7	8.4	6.25	11.1	10.0	15.8	6
CGG	10.7	1.1	0	4.4	0	0	2
AGA	17.4	43.7	39.5	28.8	45.0	47.4	44
AGG	21.7	26.0	27.0	22.2	25.0	10.5	22
Gly GGT	27.1	35.8	23.0	37.5	19.0	33.3	34.6
GGC	39.5	21.7	33.3	12.5	19.0	13.3	3.8
GGA	19.7	34.0	33.3	43.8	47.6	46.6	53.8
GGG	13.6	8.5	10.2	6.2	14.3	6.6	7.7

* only 1 cys present
** only 2 cys present

Pro CCG, Thr ACG, Ala GCG and also the 4 codons for Arg beginning CG. However, interestingly, *Leg* J and *Vic* J both show higher levels of usage for Arg CGT and CGA than do any other of the other pea storage protein genes (See below for further discussion). This general avoidance of codons containing CG dinucleotides is presumably linked with the well established phenomenon of CG suppression, observed in both animal and plant genes in general, (Bird., 1980, McClelland and Ivarie., 1982, and McClelland, 1983.) and also specifically in pea seed storage proteins (Waterhouse 1985).

Secondly, it has been observed that plant genes only tend to avoid TA in positions 2 and 3 of the codon, as opposed to animal genes which show avoidance of AA, AT, TA and TT in these positions (Lycett et al., 1983b). Some of the pea seed genes don't even seem to show avoidance of TA - e.g. *Vic* β and *Vic* J for Val GTA, and Ile ATA. This seems to strongly suggest that any constraint which might be acting on animal genes causing them to avoid usage of the above dinucleotides, is not apparently active in plant genes.

Thirdly, several observations can be made by comparison of patterns only in the pea seed genes.

i) *Leg* A and *Leg* J employ opposite patterns of usage for the two synonymous codons for Tyr

ii) Both *Vic* J and CVA demonstrate extremely high levels of preferrence for CAA rather than CAG to code for Gln.

iii) *Leg* J is the only gene to show preference for GAC rather than GAT in coding for Asp.

iv) Particularly CVA, but also *Leg* J and *Vic* J show a much lower preference for Val GGC than that observed for *Leg* A (and also to a lesser extant, *Vic* B).

v) Finally, it should be pointed out that patterns observed for Cys are not significant due to the very small number of residues contributing to these figures.

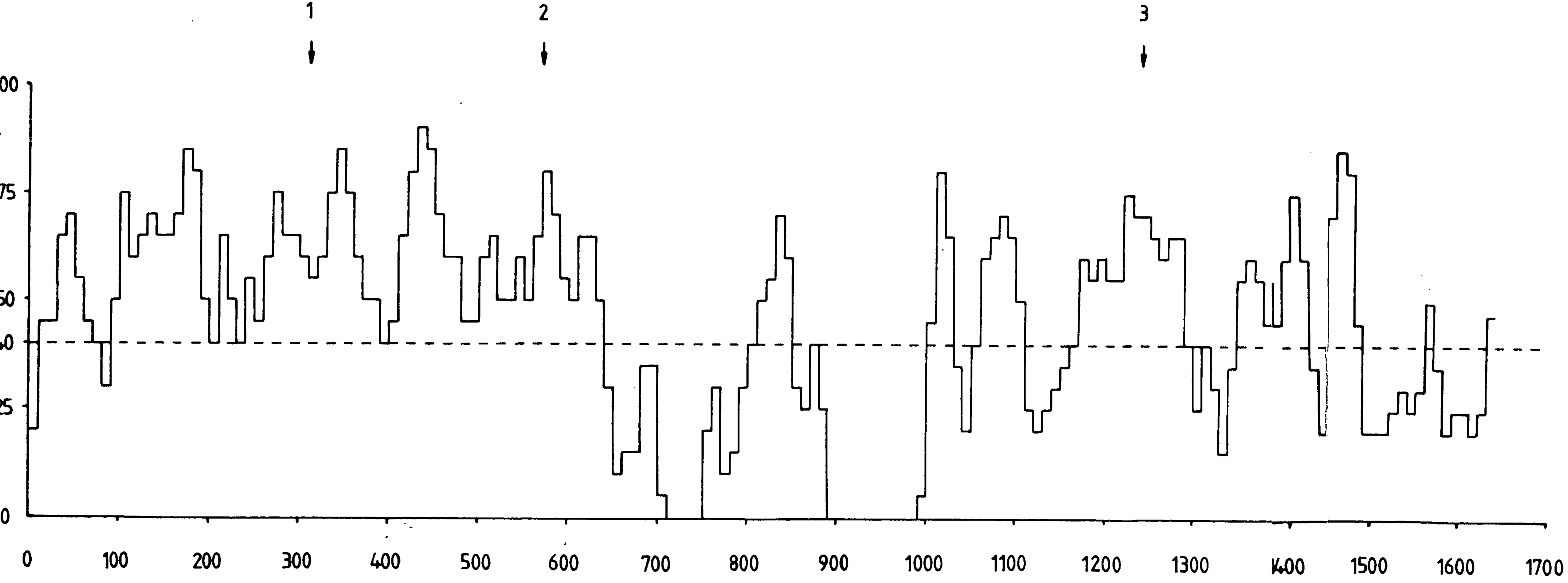
Additionally, results from this table can be compared with codon usage figures for *Gma* β and *Pvu* β (Doyle et al., 1986). *Gma* β and *Pvu* β show preference for AAC rather than AAT for Asn, with the opposite being the case for *Vic* B, *Vic* J and CVA. Also *Gma* β and *Pvu* β favour GTT and GTG over GTC and GTA for Val, whereas GTA is strongly favoured in *Vic* B, *Vic* J and CVA, with GTC also being of surprisingly high prefer ence in *Vic* J.

The significance of these observations is debatable. They could imply that individual patterns of codon usage have evolved since divergence of the genes from their common ancestor, to explain the different patterns observed between the more closely related *Gma* β and *Pvu* β as opposed to the pea vicilin genes (see section 4.2.2.). However, since differences in usage pattern are also observed between just the pea seed genes, one might also conclude that the levels of difference observed are not significant in terms of regulating levels of expression via translation, as opposed to bacterial genes, where achieving maximum translation rates is extremely important, and therefore the relationship between codon usage and levels of various tRNAs is maintained to a much greater degree (Bulmer, 1987). Indeed the situation in storage protein genes may be similar to that discussed earlier in mammalian β globin genes (see above). A possible way of elucidating the situation might be use of synthetic olignonucleotide insertions containing normally avoided codons, followed by analysis of the translational efficiency of the resulting protein. Such an approach proved successful for *E.coli* , where insertion of such oligonucleotides into a highly expressed gene caused a reduction in its translation rate (Robinson et al., 1984).

Finally, it is of interest to note the different patterns of usage observed between *Leg* A, coding for a 'major' legumin polypeptide species, and *Leg* J, encoding a 'minor' legumin polypeptide. The two genes represent the two legumin gene sub-families, and although the results are not given in Table 14, *Leg* B, C and D show virtually identical patterns of usage to that of *Leg* A, while *Leg* K shares a similar pattern to *Leg* J. From observations made above on usage differences between *Leg* A and *Leg* J, one thoery might be that the J and K genes occupy a 'mopping up' role in storage protein synthesis i.e. *Leg* A type genes would be responsible for the bulk of protein synthesis, and that any residual abundant tRNA species left over from their translation, along with those less abundant tRNAs perhaps not favoured by the *Leg* A type genes, would then, rather than be wasted, get used in translation of *Leg* J and *Leg* K. This is obviously a highly speculative theory, and much confirmatory work would be needed to support it.

4.3.2 Codon Adaptation Index Values for Pea Seed Storage Protein Genes

A possible method of analysing the potential for the involvement of codon usage in gene expression is the notion of the Codon Adaption


1
2
3
00
75
50
40
25
0
0
100
200
300
400
500
600
700
800
900
1000
1100
1200
1300
1400
1500
1600
1700

Index or CAI (Sharp and Wen-Hsuing, 1987). A method for calculating the CAI of a particular gene is given in section 2.13.1. Basically, the theory behind the method is that by using patterns of codon usage from a set of highly expressed genes as a reference, then a figure for relative suitability of each codon from a synonymous set can be calculated, and using the method outlined, an overall figure representing the pattern of codon bias, the CAI, can be calculated. So far the method has only been performed for *E.coli* and *S.cerevisiae* where a clear correlation between the degree of codon bias and level of gene expression has already been established (Gouy and Gautier, 1982., and Bennetzen and Hall, 1982). Such a relationship is yet to be confirmed in any eukaryotic gene system, let alone specifically in legume seed storage proteins, so with this in mind, the calculation of CAI values for pea seed storage protein genes using reference values obtained from codon usage figures from all available legume storage protein genes (see Table 4, section 2.13.1) was obviously highly speculative. The resulting values are given below in Table 15.

Table 15 : CAI values of pea seed storage proteins.

Gene	CAI value
Leg A	0.6519
Leg J	0.6313
Leg K	0.6129
Vic B	0.6478
Vic J	0.6343
CVA	0.6377

N.B. Values for *Leg* B, C and D were virtually identical to that of *Leg* A and are thus not given.

The application of this method to any eukaryotic system will always be fraught with difficulties, not least in selecting a suitable set of highly expressed reference genes. In this case, it was considered inadequate to use pea seed storage proteins only as the reference set, since not only would such a reference set be too small, but also genes such as those of the *Leg* A type would obviously be reflected as having a more favourable CAI than would Leg J or K, since there would be more *Leg* A type genes in the reference set. Consequently, a set of R.S.C.U. values was constructed using data from all available legume seed storage protein genes (see Table 4). Such an approach was considered at least partly justifiable from the observations made in higher eukaryotes (sheep haemoglobins, Litt and Kabat, 1972., rabbit and bovine liver and brain, Hatfield et al., 1979 and silkglands of *Bombyx mori* , Garel 1974) in that changes in tRNA profiles can be observed to occur in a tissue - specific manner (Bennetnen 1982), to the extent that the resulting isoacceptor tRNA distribution matches the codon usage bias of abundant tissue-specific mRNAs. Thus, a putative assumption was made that, since legumes have evolved from a common ancestor (see section 4.2) it is possible that some degree of conservation (already observed to have been operating at the nucleic acid and amino acid sequence levels) may also have operated on the overall levels of isoaccepting tRNAs in legume embryo tissue. Assuming this to be so, then calculation of CAI values in the way stipulated above could potentially reveal, in a similar way to that observed in bacteria and also a lower eukaryote, yeast (Sharp et al., 1986, Sharp and Wen-Hsuing, 1987), the relative levels of expression of each of the pea seed storage protein genes. (i.e. Those genes using generally preferred codons most often would achieve the highest CAI values, and would be expressed at higher levels due to the greater abundance of the tRNA species that their codon usage pattern demands. Correspondingly, genes showing a preference for codons not generally in favour amongst other legume genes would consequently achieve a lower CAI value, and be expressed at a lower level due to lower availability of the required tRNA species).

Bearing this in mind, the findings given in Table 15 appear quite interesting. Perhaps most significant of all is the difference in CAI value observed between *Leg* A and *Leg* J, suggesting *Leg* A to be expressed to a higher degree then *Leg* J, giving a potential numerical confirmation of a similar suggestion made earlier (section 4.3.1). It is also interesting to note that *Leg* K is assigned a CAI value considerably lower than that of *Leg* J, despite being a member of the same sub-gene family. This could even be interpreted as a further extension of the 'mopping up' theory (section 4.3.1), with *Leg* K playing a secondary role to that of *Leg* J in such a potential system of translation.

Figures for the vicilin genes are less revealing, and since both *Vic* B and *Vic* J are no longer thought to be expressed anyway, for reasons mentioned earlier, these figures will not be discussed further.

Obviously, before any findings of this kind can be routinely accepted, further data on the efficiency of translation *in vivo* , of the genes in question, would be needed. However, should it one day be confirmed that levels of expression of a gene can indeed be modulated via tRNA levels in relation to patterns of codon usage, such a finding could be significant in explaining observed levels of expression of foreign genes in heterologous systems, where levels of tRNA species may not be ideal for efficient translation of the foreign gene.

4.4 Gene Hybridisation Studies of Several Different Pea Lines

The original objective of this study was to investigate whether any detectable gene duplication events could be observed in a comparison of gene copy numbers of specific storage protein genes from a range of pea lines of varying age. As will be discussed below, this objective was at least partly achieved, although, perhaps not quite showing the results expected. However, it did not prove possible to accompany these studies with any conclusions on the relatedness of gene copy number to level of gene product in any of the pea lines. In order to investigate this area further, detailed analysis on the amounts of 11S and 7S storage protein deposition in each of the lines would be needed.

Several observations can be made on this work from the autoradiographs and Table 5 presented in section 3.

Firstly, allowing for one or two possible anomalies in the blotting or hydridisation procedure (e.g. *Leg* A hybridised to Mangetout DNA cut with Eco R1 appeared to give only 3 very feint bands, whereas the same probe hybridised to the same DNA cut with Hind III gave a much stronger and more complex band pattern. The former result was thought to be anomalous). Additionally, consideration was given to the accuracy with which the concentration of genomic DNA in each pea line extract was determined (using the spectrophotometric method outlined in section 2.12.). Assuming that such a method was inevitably prone to a certain degree of inaccuracy, consequently, some slight difference in amounts of DNA loaded onto the gels from each line was anticipated. However, anomalies apart, it seemed clear from an overall assessment of the results of each blot, that levels of gene copy number for each of the genes studied did not appear to differ to any significant degree i.e. each pea line appeared to possess equal copy numbers for each of the genes. This conclusion is obviously only of a tentative nature, since comparisons between lines became much more difficult in cases where band patterns differed for a given probe between each of the lines. Indeed the conclusion might be consolidated by laser scanning desitometry in order to assess and compare overall band intensities for each gene between the lines.

The only case where it was considered that a duplication event might have occured was in the case of *Vic* B hybridised to EcoR1 digested DNA samples. Here it appeared that a band of 4.7 kb (representing *Vic* B itself, see section 4.1.3.3.) was approximately twice as intense in Feltham First, 1263, JI81, 851, 807 and 809 as it was in 1552, which was in turn approximately twice as intense as this band in 808, implying a possible double duplication event. However, these findings are extremely tentative, and the observed differences in intensities of this band could also be due to mutations at the Eco R1 sites around this region of the genome, giving consequent differences in band hybridisation patterns.

This apart, the findings that gene copy numbers appear similar for each gene in each of the pea lines is consistent with a similar observation made on several pea genotypes by Domoney and Casey, 1985. Here DNA from each genotype was probed with representative genes from both the 11S and 7S multigene families and gene copy numbers appeared similar in each genotype. Indeed these studies were linked to an analysis of 11S and 7S storage protein deposition which interestingly was shown to differ for each genotype, implying that product levels appear to be regulated by transcriptional/translational processes, rather than simply by gene number. Such a conculsion cannot be drawn from the data presented here for the reasons mentioned above, but, assuming that levels of protein deposition do vary considerably between each pea line, based on the great difference in average seed diameter observed between the lines, (see Table 16) a similar conclusion might be drawn - i.e. that regulation occurs at the transcription/translation level, and that gene copy number has no direct influence on levels of protein deposition.

Table 16 Average Sizes (diameters) of Seeds from the pea lines studied

Line	Average Diameter*(mm)	Standard Deviation
Feltham First	8.07	1.12
JI 81	5.1	0.7
851	6.34	1.52
807	4.6	0.56
808	5.12	0.73
809[2]	3.89	0.40
1263	5.87	0.71
1552	7.39	0.91
Mangetout	7.06	0.82

* Sample size of 6, taking 5 random diameter measurements on each using a micrometer screw gauge.

A second interesting observation made from these results is the considerably greater degree of conservation in hybridisation patterns shown on Hind III digested samples as compared to those digested with Eco RI. Such a finding would seem to be adequately explained by the fact that Hind III sites tend to occur in or near the coding sequences of the genes examined, and are thus subjected to more evolutionary constraint, as opposed to Eco RI sites, which tend to lie outside the coding regions, where constraint acts to a much lesser extent, and are thus more prone to the continual process of mutation.

A final observation from this work was made when comparing the results of hybridising *Leg* K plus approximately 2kb of its 3' noncoding sequence to samples digested with Eco RI, and of hybridising purely *Leg* K coding sequence to samples digested in the same way. From the findings it was clear that the 2Kb of sequence 3' to *Leg* K contains sequence(s) that are very highly repeated throughout not only the Feltham First genome, but also the genomes of all the other pea lines, implying an early origin for the repeated sequence(s) detected here.

APPENDIX

CHARACTERISTICS OF THE PEA LINES USED FOR DNA EXTRACTIONS

Genetic markers

807	Kp, D < ma, M
808	pafl, up, D < ma, fl, di, Pl, u
809	kp, td, D < ma, M, F, Fs
851	pa, vim, b, K, dt, pr, wb, tl < w, st, td, D < co, fl, pro, le, di mifo, s, oh, cor, F, Fs, i, r
1256	Cit, Cm, kp, pafl, dt, pr, up, Ser, d, fru, pro, Astr, wp, Pl, M, Umb, F, Fs
1263	a, Br, dt, Tra
1552	/r
JI	Not known
Mangetout	Not known

Specification

807	L K CO	P. arvense
808	L K CO	P. abyssinicum
809	L K CO	P. asiaticum
851	L K MU	P. procumbens
1256	L K CO	P. fulvum
1263	L K CV	P. weitor
1552	E K CO	Ubamer P. speciosum

L	=	Line
E	=	wildform collections, primitive varieties, landraces
K	=	normal constant material (with respect to mode of maintenance)
MU	=	mutant
Cr	=	cultivar
CO	=	collection

Genesymbols of Pisum

Kp	Keel with AB CV coloured with anthocyanin	Dom.
D < ma	Disappearance of axil colour : two spots	Rec.
M	Brown marbling of testa	Dom.
pafl	Small flower size	Inc. Dom.
up	Only one pair of leaflets	Rec.
fl	No flecking on leaflets and stipules	Inc. Dom.
di	More or less deep impressions in tests and cotyledons	Rec.
Pl	Black in hilum colour	Dom.
u	Testa uniform violet	Rec.
td	No dentation or occasional single tooth on leaflet	Inc. Dom.
F	Violet spots on testa	Dom.
Fs	Violet spots bleached with b	Dom.
pa	Medium green colour of foliage, pods and immature seeds	Rec.
Vim	Medium dark green foliage and pod colour	Rec.
b	Pink flowers	Rec.
k	Wings reduced, adpressed to keel	Rec.
dt	Shortens pedunele length from axil to 1st flower	Rec.
pr	Shortens inflorescence	Rec.
wb	Pods waxless, stipules on both sides and under leaflets waxless	Rec.
tl < v	Tendrils converted to leaves	Inc. Dom.
st	Stipules lanceolate, slightly bent, surface reduced by approx. 80%	Rec.
D < c	Disappearance of axil colour - a single ring	Rec.
pro	Basal branches growing at 45° angle	Rec.
le	Short internodes, zig zag pattern, later flowering and shorter roots. Leaf and pod colours darker.	Rec.

mifo	Close set small and ahsllow impressions on testa	Rec.
s	Seeds glued together in pod	Rec.
oh	Testa reddish brown. Dark spot near hilum	Rec.
cor	Ochraceous coloured hilum region	Inc. Dom.
i	Cotyledons green	Rec.
r	Cotyledons wrinkled. Starch grains compound. Amylose sugar content and water uptake higher.	Rec.
Cit	Flower colour citrus yellow to cream	Dom.
Cm	Flower colour coral-rose	Dom.
Ser	Serratus dentation, saw toothed	Inc. Dom.
d	Disappearance of axil colour	Rec.
fru	Increased number of basal branches	Rec.
Astr	Pods with purple-violet, short, longitudinal stripes	Dom.
wp	Pods waxless	Inc. Dom.
Umb	Umber coloured testa	Dom.
a	Absence of anthocyanin production	Rec.
Br	With bracts of varying size	Dom.
Tra	Tragacanth on inside of testa, visible from outside as only spot	Inc. Dom.

CHAPTER 5 - BIBLIOGRAPHY

Argos, P., Narayana, S.V.L., and Nielsen, N.C. 1985; Structural similarity between legumin and vicilin storage proteins from legumes. EMBO J. 4 1111-1117.

Axel, R., Feigelson, P., and Schutz, G. 1976; Analysis of the complexity and diversity of mRNA from chicken oviduct and liver. Cell 7 247-254.

Badenoch-Jones, J., Spencer, D., Higgins, T.J.V., and Millerd, A. 1981; The role of glycosylation in storage protein synthesis in developing pea seeds. Planta 153 201-209.

Banerji, J., Rusconi, S., and Schaffner, W. 1981; Expression of a genomic segment of Simian virus 40 DNA. Cell 27 299-304.

Barrier, P.A., Jeffreys, A.J., and Scott, A.F. 1981; Evolution of the β-globin gene cluster in man and primates. J. Mol. Biol. 149 319-336.

Barton, K.A., Thompson, J.F., Madison, J.T., Rosenthal, R., Jarvis, N.P., and Beachy, R.N. 1982; The Biosynthesis and processing of high molecular weight precursors of soybean glycinin subunits. J. Biol. Chem. 257 6089-6095.

Bassuner, R., Huth, A., Manteuffel, R., and Rapoport, T.A. 1983; Secretion of Plant Storage Globulin polypeptides by Xenopus laevis oocytes. Eur. J. Bioch. 133 321-326.

Baumlein, H., Wobus, U., Pustell, J. 1986; The legumin gene family, structure of a B type gene of V. faba and a possible legumin gene specific regulatory element. N.A.R., 14 2707-2720.

Bedbrook, J.R., Jones, J., O'Dell, M., Thompson, R., and Flavell, R.B. 1980a; A molecular description of telomeric heterochromatin in Secale species. Cell 19 545-560.

Bedbrook, J.R., O'Dell, M., and Flavell, R.B. 1980b; Amplification of rearranged sequences in cereal plants. Nature 288 133-137.

Bennett, M.D. 1982; Nucleotypic basis of the spatial ordering of chromosomes in eukaryotes and the implications of the order for genomic evolution and phenotypic variation. In "Genome Evolution" (Dover, G.A., and Flavell, R.B., eds), Academic Press, London.

Bennetzen J.L., and Hall, B.D. 1982; Codon Selection in Yeast. J. Biol. Chem. 257 3026-3031.

Benoist, C., and Chambon, P. 1980; Deletions covering the putative promoter region of early messenger RNA species of SV40 do not abolish T (tumor)-antigen expression. P.N.A.S. 77 3865-3869.

Benoist, C., O'Hare, K., Breathnach, R., and Chambon, P. 1980; The ovalbumin gene-sequence of putative control regions. N.A.R. 8 127-142.

Bergmann, I.E., and Brawerman, G. 1977; Control of breakdown of the polyadenylation sequence in mammalian polyribosomes : role of poly(adenylic acid) -Protein interactions. Biochemistry 16 259-264.

Bird, A.P., 1980; DNA methylation and frequency of CpG in animal DNA. N.A.R. 8 1499-1504.

Birnboim, H.C., and Doly, J. 1979; A rapid alkaline extraction procedure for screening recombinant plasmid DNA. N.A.R. 7 1513-1523.

Birnsteil, M.L., Busslinger, M., and Strub, K. 1985; Transcription Termination and 3' processing, the end is in site. Cell 41 349-359.

Black, D.L., and Steitz, J.A. 1986; Pre-MRNA splicing *in vitro* requires intact U4/U6 small nuclear ribonucleoprotein. Cell 46 697-704.

Bolivar, F., Rodriguez, R.L., Green, P.J., Betlach, M.C., Heynecker, H.L., Boyer, H.W., Crosa, J.H., and Falkow, S. 1977; Construction and characterisation of new cloning vehicles. II. A multipurpose cloning system. Gene 2 95-113.

Bollini, R., and Chrispeels, M.J. 1978; Characterisation and subcellular localisation of vicilin and phytohaemaglutinin, the two major reserve proteins of *Phaseolus vulgaris* L. Planta, 142 291-298.

Bollini, R., and Vitale, A. 1981; Genetic variability in charge microheterogeneity and polypeptide composition of phaseolin, the major storage protein of *Phaseolus vulgaris* and peptide maps of its three subunits. Physiol Plant 52 96-100.

Bollini, R., Vitale, A., and Chrispeels, M.J. 1983; *In vivo* and *in vitro* processing of seed reserve protein in the Endoplasmic Reticulum : Evidence for two glycosylation steps. J Cell Biol. 96 999-1007.

Borrotto, K., and Dure, L. 1987; The globulin seed storage proteins of flowering plants are derived from two ancestral genes. Pl. Mol. Biol. 8 113-131.

Boulter, D. 1982; The composition and nutritional value of legumes in relationship to crop improvement by breeding. Proc. Nutr. Soc. 41 1-6.
* see p.237

Boutry, M., and Chua, N.H. 1985; A nuclear gene encoding the beta subunit of the mitochondrial ATP synthase in Nicotiana plumbaginofolia. EMBO J. 4 2159-2165.

Bown, D., Ellis, T.H.N., Gatehouse, J.A. 1988 (in press); A 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.

Breathnach, R., Benoist, C., O'Hare, K., Gannon, F., and Chambon, P. 1978; Ovalbumin gene: evidence for a leader sequence in mRNA and DNA sequences at the exon-intron boundaries. P.N.A.S. 75 4853-4858.

Brinegar, A.C., and Peterson, D.M. 1982; Synthesis of Oat globulin precursors. Plant Physiol. 70 1767-1769.

Bringmann, P., Appel, B., Rinke, J., Reuter, R., Theissen, H., and Luhrmann, R. 1984; Evidence for the existence of snRNAs U4 and U6 in a single ribonucleoprotein complex and for their association by intermolecular base pairing. EMBO J. 3 1357-1363.

Brisson, N., and Verma, D.S.P. 1982; Soybean leghaemoglobin gene family; normal, pseudo and truncated genes. P.N.A.S. 79 4055-4059.

Britten, R., and Kohne, D. 1968; Repeated sequences in DNA. Science 161 529-540.

Brown, D.D., Wensink, P.C., and Jordan, E. 1971; Position and characteristics of 5S DNA from X. laevis. P.N.A.S. 68 3175-3179.

Brown, J.W.S. 1986; A catalogue of splice junction and putative branch point sequences from plant introns. N.A.R. 14 9549-9559.

Brown, J.W.S., Ma, Y., Bliss, F.A., and Hall, T.C. 1981a; Genetic variation in the subunits of globulin-1 storage protein of French bean (Phaseolus vulgaris). Theor. Appl. Genet. 59 83-88.

Brown, J.W.S., Bliss, F.A., and Hall, T.C. 1981b; Linkage relationships between genes controlling seed proteins in French bean (Pheseolus vulgaris). Theor. Appl. Genet. 60 251-258.

Brown, J.W.S., Feix, G., and Frendewey, D. 1986; Accurate in vitro splicing of two pre-mRNA plant introns in a Hela Cell nuclear extract. EMBO J. 5 2749-2758.

Buckingham, M.E., and Minty, A.J. 1983; Contractile protein genes. In : Eukaryotic Genes, their structure, activity, and regulation. Butterworth Publications.

Bulmer, M. 1987; Coevolution of codon usage and transfer RNA abundance. Nature 325 728-730.

Casey, R. 1979a; 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. 1979b; Immunoaffinity chromatography as a means of purifying legumin from Pisum (Pea) seeds. Biochem. J. 177 509-520.

** see p. 237

Casey, R., Domoney, C., and Stanley, J. 1984; Convicilin mRNA from pea (Pisum sativum) has sequence homology with other legumin 7S storage protein mRNA species. Biochem. J. 224 661-666.

Catsimpoolas, N., and Ekenstam, C. 1969; Isolation of Alpha, Beta, and Gamma conglycinins. Arch. Biochem. Biophys. 129 490-497.

Catsimpoolas, N., Kenney, J.A., Meyer, E. W. 1971; The effect of thermal denaturation on the antigenicity of glycinin. Biochem. Biophys. Acta 229 451-458.

Cavener, D. R. 1987; Comparison of the consensus sequence flanking translational start sites in Drosophila and vertebrates. N.A.R. 15 1353-1361.

Chen, C. W., and Thomas, C. A. 1980; Recovery of DNA segments from Agarose gels. Anal. Biochem. 101 339-341.

Chen, J., and Varner, J.E. 1985; An extracellular matrix protein in plants: Characterisation of a genomic clone for carrot extensin. EMBO J. 4 2145-2150.

Chen, Z.L., Schuler, M.A., and Beachy, R.N. 1986; Functional analysis of regulatory elements in a plant-embryo specific gene. P.N.A.S. 83 8560-8564.

Chrispeels, M.J., Higgins, T.J.V., Craig, S., and Spencer, D. 1982a; Role of the endoplasmic reticulum in the synthesis of reserve proteins and the kinetics of their transport to protein bodies in developing pea cotyledons. J. Cell Biol. 93 5-14.

Chrispeels, M.J., Higgins, T.J.V., and Spencer, D. 1982b; Assembly of storage protein oligomers in the endoplasmic reticulum and processing of the polypeptides in the protein bodies of developing pea cotyledons. J. Cell Biol. 93 306-313.

Cronquist, A. 1981; An integrated system of classification of flowering plants. Columbia University Press, New York.

Croy, R.R.D., Derbyshire, E., Krishna, T.G., and Boulter, D. 1979; Legumin of *Pisum sativum* and *Vicia faba*. New Phytol. 83 29-35.

Croy, R.R.D., and Gatehouse, J.A. 1985; In Plant Genetic Engineering, Ed Dodds, J.H., Camb. Univ. Press, 143-268.

Croy, R.R.D., Gatehouse, J.A., Tyler, M., and Boulter, D. 1980; The purification and characterisation of a third storage protein (convicilin) from the seeds of pea (*Pisum Sativum* L.). Biochem. J. 191 509-516.

Croy, R.R.D., Gatehouse, J.A., Evans, I.M., and Boulter, D. 1980a; Characterisation of the storage protein subunits synthesised *in vitro* by polyribosomes and RNA from developing pea (*Pisum sativum* L.). Planta 148 49-56.(1.Legumin)

Croy, R.R.D., Gatehouse, J.A., Evans, I.M., and Boulter, D. 1980b; Characterisation of the storage protein subunits synthesised *in vitro* by polyribosomes and RNA from developing Pea (*Pisum sativum* L.). Planta 148 57-63.(2 Vicilin)

Croy, R.R.D., Hoque, M.S., Gatehouse, J.A., and Boulter, D. 1984; The major albumin proteins from pea (*Pisum Sativum* L.). Biochem. J. 218 795-803.

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

Czelusniak, J., Goodman, M., Hewett-Emmett, West, M.L., Venta, P.J., and Tashian, R.E. 1982; Phylogenetic origins and adaptive evolution of avian and mammalian haemoglobin genes. Nature 298 297-300.

Davey, R.A., Higgins, T.J.V., and Spencer, D. 1981; Homologies between two small subunits of vicilin from *Pisum sativum* L.). Biochem.Int. 3 595-602.

Davies, C.S., Coates, J.B., and Nielsen, N.C. 1985; Inheritance and biochemical analysis of four electrophoretic variants of β-conglycinin from soybean. Theor. Appl. Genet. 71 351-358.

Delauney, A.J. 1984; Cloning and characterisation of cDNAs encoding the major pea storage proteins, and expression of vicilin in E. Coli. PhD. Thesis, University of Durham.

Dennis, E.S., Dunsmuir, P., and Peacock, W.J. 1980a; Segmental amplification in a satellite DNA:restriction enzyme analysis of the major satellite of Macropus rufogriseus. Chromosoma 79 179-198.

Dennis, E.S., Gerlach, W.L., and Peacock, W.J. 1980b; Identical poly-pyrimidine-polypurine satellite DNAs in wheat and barley. Heredity 44 349-366.

Dennis, E.S., Gerlach, W.L., Pryor, A.J., Bennetzen, J.L., Inglis, A., Llewellyn, D., Sachs, M.M., Ferl, R.J., and Peacock, W.J. 1984; Molecular analysis of the alcohol dehydrogenase 2(Adh2) gene of maize. N.A.R. 12 3983-4000.

Derbyshire, E., Wright, D.J., and Boulter, D. 1076; Legumin and vicilin, storage proteins of legume seeds. Phytochemistry 15 3-24.

Dierks, P., Van Ooyen, A., Mantei, N., and Weissmann, C. 1981; DNA sequences preceding the rabbit β-globin gene are required for formation of mouse L cells of β-globin RNA with the correct 5'terminus. P.N.A.S. 78 1411-1415.

Docherty, K., Carroll, R.J., and Steiner, D.F. 1982; Conversion of Proinsulin to insulin; involvement of a 31,500 molecular weight thiol protease. P.N.A.S. 79 4613-4617.

Domoney, C., and Casey, R. 1983; Cloning and characterisation of complementary DNA 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, an 80,000 Mr legumin related polypeptide. Eur. J. Bioch. 139 321-327.

Domoney, C., and Casey, R. 1985; Measurement of gene number for seeds storage proteins in Pisum. N.A.R. 13 687-699.

Donahue, T.F., Davies, R.S., Lucchini, G., and Fink, G.R. 1983; A short nucleotide sequence required for regulation of HIS4 by the general control system of yeast. Cell 32 89-98.

Doyle, J.J., Schuler, M.A., Godette, W.D., Zenger, V., Beachy, R.N., and Slightom, J.L. 1986; The glycosylated seed storage proteins of Glycine max and Phaseolus vulgaris-structural homologies of genes and proteins. J. Biol. Chem. 261 9228-9238.

Dover, G.A., 1982; A role for the genome in the origin of species? In : "Mechanisms of Speciation" (Barigozzi, C., Montalenti, G., and White M.J.D., eds). Prog. in Clinical and Biological Res. Ser. v.96 A.R. Liss.

Dover, G.A., Brown, S.D.M., Coen, E.S., Dallas, J., Strachan, T., and Trick, M. 1982; The dynamics of genome evolution and species differentiation. In : "Genome Evolution" (Dover, G.A., and Flavell, R.B., eds.), Academic Press, London.

Dover, G.A., and Coen, E. 1981; Springcleaning ribosomal DNA - a model for multigene evolution? Nature 290 731-732.

Dynan, W.S., and Tjian, R. 1985; Control of eukaryotic mRNA synthesis by sequence specific DNA-binding proteins. Nature 316 774-78.

Early, P., Rogers, J., Davis, M., Calame, K., Bond, M., Wall, R., and Hood, L. 1980; Two mRNAs can be produced from a single immunoglobulin µ-gene by alternative RNA-processing pathways. Cell 20 313-319.

Efstratiadis, A., Posakony, J.W., Maniatis, T., Lawn, R.M., O'Connell, C., Spritz, R.A., DeRiel, J.K., Forget, B.G., Weissmann, S.M., Slightom, J.L, Blechl, A.E., Smithies, O., Baralle, F.E., Shoulders, C.C., and Proudfoot, N.J. 1980; The structure and evolution of the human β -glogin gene family. Cell 21 656-668.

Ellis, T.H.N., Domoney, C., Castleton, J., Cleary, W., and Davies, D.R. 1986; Vicilin genes of Pisum. M.G.G. 205 164-169.

Evans, I.M., Croy, R.R.D., Hutchinson, P., Boulter, D., Payne, P.I., and Gordon, M.E. 1979; Cell free synthesis of some storage protein subunits by polyribosomes and RNA isolated from developing seeds of pea (Pisum Sativum L.). Planta 144 455-63.

Ferguson, A. 1980; Biochemical systematics and evolution. Halstead Press.

Fischer, R.L., and Goldberg, R.B. 1982; Structure and flanking regions of soybean seed protein genes. Cell 29 651-660.

Fitch, W.M., 1976; The estimate of total nucleotide substitutions from pairwise differences in biased. Phil. Trans. R. Soc. Lond. B. 312 317-325.

Fitzgerald, M., and Schenk, T. 1981; The sequence 5'-AAUAAA-3' forms part of the recognitiion site for polyadenylation of late SV40 mRNAs. Cell 24 251-260.

Flavell, R.B., 1982; Sequence amplification, deletion, and rearrangement; Major sources of variation during species diveregence. In : "Genome Evolution" (Dover, G.A., and Flavell, R.B., eds.), Academic Press, London.

Flavell, R.B., 1980; The molecular characterisation and organisation of plant chromosomal DNA sequences. Ann. Rev. Pl. Physiol. 31 569-596.

Flavell, R.B., O'Dell, M., and Hutchinson, J. 1981; Nucleotide sequence organisation in plant chromosomes and evidence of sequence translocation during evolution. C.S.H. Symp. Quant. Biol. 45 501-508.

Focus 6(3), 1984; p.13.

Frendewey, D., and Keller, W. 1985; Stepwise assembly of a pre-mRNA splicing complex requires U-snRNAs and specific intron sequences. Cell 42 355-367.

Fukazawa, C., Momma, T., Hirano, H., Harada, K., and Udaka, K. 1985; Glycinin A_3B_4 mRNA-cloning and sequencing of double-stranded cDNA complementary to a soybean storage protein. J. Biol. Chem. 260 6234-6239.

Furiuchi, Y., La Fiandra, A., and Shatkin, A.J. 1977; 5' terminal structure and mRNA stability. Nature 266 235-239.

Garel, J.P. 1974; Functional adaptation of tRNA population. J. Theor. Biol. 43 211-225.

Gatehouse, J.A., Bown, D., Gilroy, J., Levasseur, M.D., Castleton, J., and Ellis, T.H.N. 1988; Two genes encoding 'minor' legumin polypeptides in pea (Pisum sativum L). Biochem. J. 250 15-24.

Gatehouse, J.A., Bown, D., Evans, I.M., Gatehouse, L.N., Jobes, D., Preston, P., and Croy, R.R.D. 1987; The sequence of the seed lectin gene from pea (Pisum sativum L.). N.A.R. 15 7642.

Gatehouse, J.A., Croy, R.R.D., Morton, H., Tyler, M., and Boulter, D. 1981; Characterisation and subunit structures of the vicilin storage proteins of Pea (Pisum sativum L.). Eur. J. Biochem. 118 627-633.

Gatehouse, J.A., Evans, I.M., Bown, D., Croy, R.R.D., and Boulter, D. 1982a; Control of storage-protein synthesis during seed development in pea. Biochem. J. 208 119-127.

Gatehouse, J.A., Lycett, G.W., Croy, R.R.D., and Boulter, D. 1982b; The post-translational proteolysis of the subunits of vicilins from pea (Pisum sativum L.). Biochem. J. 207 629-652.

Gatehouse, J.A., Evans, I.M., Croy, R.R.D., and Boulter, D. 1986; Differential expression of genes during legume seed development. Phil. Trans. R. Soc. Lond. B. 314 367-384.

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

Gatehouse, L.N., 1986; M.Sc. Thesis. University of Durham.

Gates, M.A., 1985; Simpler DNA sequence representations. Nature 316 219.

Gayler, K.R., and Sykes, G.E. 1981; β -conglycinins in developing soybean seeds. Plant Physiol. 67 958-61.

Gerke, V., and Steitz, J.A. 1986; A protein associated with small nuclear ribonucleoprotein particles recognisens the 3' splice site of premessenger RNA. Cell 47 973-984.

Gerlach, W.L., and Peacock, W.J. 1980; Chromosomal locations of highly repeated DNA sequences in wheat. Heredity 44 269-276.

Gil, A., and Proudfoot, N.J. 1982; A sequence downstream of AAUAAA is required for rabbit β-globin and mRNA 3'-end formation. Nature 312 473-474.

Gilbert, W. 1985; Genes in pieces re-visited. Science 228 823-824.

Gilroy, J., Wright, D.J., and Boulter, D. 1979; Homology of basic subunits of Legumin from Glycine max and Vicia faba. Phytochem. 18 315-316.

Goldberg, R.B., Hoschek, G., Ditta, G.S., and Breidenbach, R.W. 1981; Abundance, diversity and regulation of mRNA sequence sets in soybean embryogenesis. Dev. Biol. 83 201-217.

Goldberg, R.B., Fischer, R.L., Harada, J.J., Jofuku, D., and Okamuro, J.K. 1983; Structure and Function of Plant Genomes, (eds Ciferri, O., and Dure, L.), Plenum Press, New York, 37-45.

Goldschmidt-Clermont, M., and Rahire, M. 1986; Sequence, evolution and differential expression of two clones encoding variant small subunits of Ribulose Bisphosphate carboxylase/oxygenase in *Chlamydamonas reinhardtii*. J. Mol. Biol. 191 421-432.

Goodman, M., Koop, B.F., Czelusniak, J., Weiss, M.L., and Slightom, J.L. 1974; The η -globin gene - its long evolutionary history in the β -globin gene family of mammals. J. Mol. Biol. 180 803-823.

Goodman, M. 1973; The chronicle of primate phylogeny contained in proteins. Symp. Zool. Soc. London 33 339-375.

Gouy, M., and Gautier, C. 1982; Codon usage in bacteria: correlation with gene expressivity. N.A.R. 10 7055-7074.

Grabowski, P.J., Seiler, S.R., and Sharp, P.A. 1985; A multicomponent complex is involved in the splicing of messenger RNA precursors. Cell 42 345-353.

Graham, D.E. 1978; The isolation of high molecular weight DNA from whole organisms of large tissue masses. Anal. Bioch. 85 609-613.

Grantham, R., Gautier, C., Gouy, M., Jacobzone, M., and Mercier, R. 1981; Codon catalogue usage is a genome strategy modulated for gene expressivity. N.A.R. 9 r43-r74.

Green, M.R., Maniatis, T, and Melton, D.A. 1983; Human β -globin pre-mRNA synthesised *in vitro* in accurately spliced in Xenopus oocyte nuclei. Cell 32 681-694.

Grohmann, K., Amalric, F., Crews, S., and Attardi, G. 1978; Failure to detect 'cap' structures in mitochondrial DNA-coded poly(A)-containing RNA from Hela cells. N.A.R. 5 637-651.

Grosschedl, R., Wasylyk, B., Chambon, P., and Birnsteil, M.L. 1981; Point mutation in the TATA box curtails expression of sea urchin H2A histone gene *in vivo*. Nature 294 178-180.

Grosveld, G.C., Shewmaker, C.K., Jat, P., and Flavell, R.A. 1981; Localisation of DNA sequences necessary for transcription of the rabbit β-globin gene in vitro. Cell 25 215-226.

Gruss, P., Dhar, R., and Khoury, G. 1981; SV 40 tandem repeated sequences as an element of the early promoter. P.N.A.S. 78 943-947.

Gruss, P. 1984; Magic enhancers. DNA 3 1-5.

Hall, T.C., Ma, Y., Buchbinder, B.U., Pyne, J.W., Sun, S.M., and Bliss, F.A. 1978; Messenger RNA for Gl protein of French bean seeds: Cell-free translation and product characterisation. P.N.A.S. 75 3196-3200.

Hall, T.C., McLeester, R.C., and Bliss, F.A. 1977; Equal expression of the material and paternal alleles for the polypeptide subunits of the major storage protein of the bean Phaseolus vulgaris L. Plant Physiol. 59 1122-1124.

Hall, T.C., Slightom, J.L., Ersland, D.R., Murray, M.G. Hoffman, L.M., Adang, M.J., Brown, J.W.S., Ma, Y., Matthews, J.A., Cramer, J.H., Barker, R.F., Sutton, D.W., and Kemp, J.D. 1983; In : Structure and function of plant genomes (eds. Ciferri, O, and Dove, L, III). Plenum Publishing Corp, New York, 123-142.

Hashimoto, C., and Steitz, J.A. 1984; U4 and U6 RNAs coexist in a single small nuclear ribonucleoprotein particle. N.A.R. 12 3283-3293.

Hatfield, D., Matthews, C.R., and Rice, M. 1979; Amino acyl-transfer RNA populations in mammalian cells - chromatographic profiles and patterns of codon recognition. Bioch. Biophys. Acta 564 414-423.

Hearing, P., and Shenk, T. 1983; The adenovirus type 5 EIA transcriptional control region contains a duplicated enhancer element. Cell 33 695-703.

Heidecker, G., and Messing, J. 1986; Structural analysis of plant genes. Ann. Rev. Pl. Phys. 37 439-466.

Henning, W. 1966; Phylogenetic systematics. Univ. of Illinois Press, Urbana.

Higgins, T.J.V. 1984; Synthesis and regulation of major proteins in seeds. Ann. Rev. Pl. Physiol. 35 191-221.

Horstmann, C. 1983; Specific subunit pairs of legumin from Vicia faba. Phytochemistry 22 1861-1866.

Hu, N-T., and Messing, J. 1982; The making of strand-specific M13 probes. Gene 17 271-277.

Hu, N-T., Peifer, M.A., Heidecker, G., Messing, J., and Rubenstein, T. 1982; Primary structure of a zein genomic clone. EMBO J 1 1337-1342.

Igo-Kemenes, T., Horz, W., and Zachau, H.G. 1982; Chromatin. Ann. Rev. Biochem. 51 89-121.

Ikemura, T. 1985; Codon usage and transfer-RNA content in unicellular and multicellular organisms. Mol. Biol. Evol. 2 13-34.

Jeffreys, A.J. 1982; Evolution of globin genes. In : Genome Evolution, (eds. Dover, G.A., and Flavell, R.B. Academic Press, London.

Jeffreys, A.J., and Harris, S. 1982; Processes of gene duplication. Nature 296 9-10.

Jacq, C., Miller, J.R., and Brownlee, G.G. 1977; A pseudogene structure in 5S DNA of Xenopus laevis. Cell 12 109-120.

Jukes, T.H., and Cantor, C.R., 1969; In : Mammalian protein metabolism, (eds. Munro, H.N., and Allison, J.B.,) Academic Press, New York, 21-132.

Jukes, T.H., and King J.L. 1979; Evolutionary nucleotide replacements in DNA. Nature 281 605-606.

John, B, and Miklos, G.L.G. 1979; Functional aspects of heterochromatin and satellite DNA. Int. Rev. Cytol. 58 1-114.

Jones, R.N., and Brown, L.M. 1976; Chromosome evolution and DNA variation in Crepis. Heredity 36 91-104.

Kafatos, F.C., Efsratiadis, D., Forget, B.G., and Weissmann, S.M. 1977; Molecular evolution of human and rabbit β -globin mRNAs. P.N.A.S. 74 5618-5622.

Katinakis, P., and Verma, D.P.S. 1985; Nodulin -24 gene of soybean codes for a peptide of the peribacteroid membrane and was generated by tandem duplication of a sequence resembling an insertion element. P.N.A.S. 82 4157-4161.

Kaulen, H., Schell, J., and Kreuzaler, F. 1986; Light-induced expression of the chimaeric chalcone synthase-NPT II gene in tobacco cells. EMBO J. 5 1-8.

Keil, M., Sanchez-Serrano, J., Schell, J., and Willmitzer, L. 1986; Primary structure of a proteinase inhibitor II gene from potato (Solanum tuberosum). N.A.R. 14 5641-5650.

Keller, E.B. and Noon, W.A. 1984; Intron splicing: A conserved internal signal in introns of animal pre-mRNAs. P.N.A.S. 81 7417-7420.

Khoury, G., and Gruss, P. 1983; Enhancer elements. Cell 33 313-314.

Kiss, T., Antal, M., and Solymosy, F. 1987; Plant small nuclear RNAs II. U6RNA and a 4.5S,-like RNA are present in plant nuclei. N.A.R. 15 543-560.

Kitamura, K., Toyokawa, T., and Harada, K. 1980; Polymorphism of Glycinin in soybean seeds. Phytochem. 19 1841-1843.

Klosgen, W.B., Gierl, A., Schwarz-Sommer, S, and Saedler, H. 1986; Molecular analysis of the waxy locus of Zea mays. M.G.G. 203 237-244.

Konarska, M.M., Padgett, R.A., and Sharp, P.A. 1984; Recognition of cap structure in splicing in vitro of mRNA precursors. Cell 38 731-736.

Kornberg, R.D. 1977; Structure of Chromatin. Ann. Rev. Biochem. 46 931-954.

Kornberg, R.D. 1981; The location of nucleosomes in chromatin: specific or statistical? Nature 292 579-580.

Kornberg, R.D. 1974; Chromatin structure: A repeating unit of histones and DNA. Science 184 868-871.

Kozak, M. 1981; Possible role of flanking nucleotides in recognition of the AUG initiator codon by eukaryotic ribosomes. N.A.R. 9 5233-5252.

Kozak, M. 1984; Compilation and analysis of sequences upstream from the translational start site in eukaryotic mRNAs. N.A.R. 12 857-872.

Kreis, M., Williamson, M.S. Forde, J., Schmutz, D., Clark, J., Buzton, B., Pywell, J., Morris, C., Henderson, J., Harris, N., Shewry, P.R., Forde, B.G., and Miflin, B.J. 1986; Differential gene expression in the developing barley endosperm. Phil. Trans. R. Soc. Lond. B. 314 355-365.

Kridl, C.J., Vieira, J., Rubsenstein, I., and Messing, J. 1982; Nucleotide sequence analysis of a zein genomic clone with a short open reading frame. Gene 28 113-118.

Krishna, T.G., Croy, R.R.D., and Boulter, D. 1979; Heterogeneity of subunit composition of the legumin of Pisum Sativum L. Phytochem. 18 1879-1880.

Lacy, E., and Maniatis, T. 1980; The nucleotide sequence of a rabbit β-globin pseudogene. Cell 21 545-553.

Langridge, P., and Feix, G. 1983; A zein gene of maize is transcribed from two widely separated promoter regions. Cell 34 1015-1022.

Laskey, R.A., and Mills, A.D. 1977; Enhanced autoradiographic detection of ^{32}P and ^{125}I, using intensifying screens and hypersensitised film. FEBS Letts 82 314-316.

Lerner, M.R., Boyle, J.A., Mount, S.M., Wolin, S.L., and Steitz, J.A. 1980; Are snRNPs involved in splicing? Nature 283 220-224.

Lewis, B. 1983; Genes, John Wiley and Sons.

Li, W-H. 1983; Evolution of duplicated genes and pseudogenes. Evolution of Genes and Proteins, (Eds Nei, M, and Koehn, R.K.), Sinauer Associates.

Litt, M., and Kabat, D. 1972; Studies of transfer ribonucleic acids and of haemoglobin synthesis of sheep reticulocytes. J. Biol. Chem. 247 6659-6664.

Little, P.F.R. 1982; Globin Pseudogenes. Cell 28 683-684.

Loomis, W.F., and Gilpin, M.E. 1986; Multigene families and vestigial sequences. P.N.A.S. 83 2143-2147.

Lueders, K., Leder, A., Leder, P., and Kuff, E. 1982; Association between a transposed α-globin pseudogene and retrovirus-like elements in the BALB/c mouse genome. Nature 295 426.

Lutcke, H.A., Chow, K.C., Mickel, F.S., Moss, K.A., Kem, H.F., and Scheele, G .A. 1987; Selection of AUG initation codons differs in plants and animals. EMBO J. 6 43-48.

Lycett, G.W., Croy, R.R.D., Shirsat, A.H., and Boulter, D. 1984; The complete nucleotide sequence of a legumin gene from Pea (Pisum sativum L.). N.A.R. 12 4493-4505.

Lycett, G.W., Delauney, A.J. Zhao, W., Gatehouse, J.A., Croy, R.R.D., and Boulter, D. 1984b; 2 complementary DNA clones coding for the legumin protein of Pisum Sativum L. contain sequence repeats. Pl. Mol. Biol. 3 91-96.

Lycett, G.W., Croy, R.R.D., Shirsat, A.H., Richards, D.M., and Boulter, D. 1985; The 5' flanking regions of three pea legumin genes : Comparison of the DNA sequences. N.A.R. 13 6733-6743.

Lycett, G.W., Delauney, A.J., Gatehouse, J.A., Gilroy, J., Croy, R.R.D., and Boulter, D. 1983a; The vicilin gene family of pea: A complete cDNA coding sequence for preprovicilin. N.A.R. 11 2367-2380.

Lycett, G.W., Delauney, A.J., and Croy, R.R.D. 1983b; Are plant genes different? FEBS Letts 153 43-46.

Ma, Y., and Bliss, F.A. 1978; Seed proteins of common bean. Crop Sci. 17 431-437.

Ma, Y., Bliss, F.A., and Hall, T.C. 1980; Peptide mapping reveals considerable sequence homology among the three polypeptide subunits of Gl storage protein from french bean seed. Plant Physiol. 66 897-902.

Maier, U-G., Brown, J.W.S., Toloczyki, C., and Feix, G. 1987; Binding of a nuclear factor to a consensus sequence in the 5' flanking region of zein genes from maize. EMBO J. 6 17-22.

Maniatis, T., Fritsh, E.M., and Sambrook, J. 1982; Molecular cloning - a laboratory manual. Cold Spring Harbor Laboratory, New York.

Manley, J.L. 1983; Accurate and specific polyadenylation of mRNA precursors in a soluble whole-cell lysate. Cell 33 595-605.

Marchionni, M., and Gilbert, W. 1986; The triosephosphate isomerase gene from maize; introns antedate the plant-animal divergence. Cell 46 133-141.

Marco, Y.A., Thanh, V.H., Tumer, N.E., Scallon, B.J., and Nielsen, N.C. 1984; Cloning and structural analysis of DNA encoding an A_2B_{1a} subunit of glycinin. J. Biol. Chem. 259 13436-13441.

Mathis, D.J., and Chambon, P. 1981; The SV40 early region TATA box is required for accurate *in vitro* initiation of transcription. Nature 290 310-315.

Matta, N.K., Gatehouse, J.A., and Boulter, D. 1981; Molecular and subunit heterogeneity of legumin of *Pisum sativum* L. (Garden Pea) - a multi-dimensional gel electrophoretic study. J. Exp. Bot. 32 1295-1307.

Mauro, V.P., Nguyen, T., Panagiotis, K., and Verma, D.P.S. 1985; Primary structure of the soybean nodulin-23 gene and potential regulatory elements in the 5'-flanking regions of nodulin and leghaemoglobin genes. N.A.R. 13 239-249.

McClelland, M. 1983; The frequency and distribution of methylatable DNA sequences in leguminous plant protein coding genes. J. Mol. Evol. 19 346-354.

McClelland, M., and Ivarie, R. 1982; Assymmetrical distribution of CpG in an 'average' mammalian gene. N.A.R. 10 7865-7877.

McGhee, J.D., and Felsenfield, G. 1980; Nucleosome structure. Ann. Rev. Biochem. 49 1115-1156.

McMaster, G.K., and Carmichael, G.G. 1977; Analysis of single- and double-stranded nucleic acids on polyacrylamide and agarose gels by using glyoxal and acridine orange. P.N.A.S. 74 4835-4838.

Messing, J., Crea, R., and Seeburg, P.H. 1981; A system for shotgun DNA sequencing. N.A.R. 9 309.

Messing, J., 1983; New M13 vectors for cloning. Methods in Enzymol. 101 20-78.

Messing, J., Geraghty, D., Heidecker, G., Hu, N-T., Kridl, J., and Rubenstein, I. 1983; Plant gene structure. In: Genetic Engineering of Plants, (Ed. Kosuge), Plenum Press.

Miflin, B.J., and Shewry, P.R. 1979; In: Seed protein improvement in cereals and grain legumes VI. Int. AEA Vienna, 137-158.

Miflin, B.J., and Shewry, P.R. 1981; In: Nitrogen and carbon metabolism (Ed Bewley, J.D.). Martinus Nijhoff/Dr W Junk, The Hague. 195-248.

Miyata, T., Yasunaga, T., Yarmawaki-Kataoka, Y., Obata, M., and Honjo, T. 1980; Nucleotide sequence divergence of mouse immunologloblin $\gamma 1$, and $\gamma 2b$ chain genes and the hypothesis of intervening sequence-mediated domain transfer. P.N.A.S. 77 2143-2147.

Momma, T., Negoro, T., Hirano, H., Matsumoto, A., Udata, K., and Fukazawa, C. 1985; Glycinin $A_5A_4B_3$ mRNA; cDNA cloning and nucleotide sequencing of soybean. Eur. J. Bioch. 149 491-496.

Momma, T., Negoro, T., Udaka, K., and Fukazawa, C. 1985; A complete cDNA coding for the sequence of glycinin A_2B_{1a} subunit precursor. FEBS Letts 188 117-22.

Moreira, M.A., Hermodson, M.A., Larkins, B.A., and Nielsen, N.C. 1981; Comparison of the primary structure of acidic polypeptides of glycinin. Arch. Biochem. Biophys. 210 633-642.

Moreira, M.A., Hermodson, M.A., Larkins, B.A., and Nielsen, N.C. 1979; Partial characterisation of the acidic and basic polypeptides of glycinin. J. Biol. Chem. 254 9921-9926.

Mount, S.M. 1982; A catalogue of splice junction sequences. N.A.R. 10 459-472.

Murray, M.G., Cuellar, R.E., and Thompson, W.F. 1978; DNA sequence organisation in the pea genome. Biochemistry, 17 5781-5790.

Murray, M.G., and Kennard, W.C. 1984; Altered chromatin conformation of the higher plant gene phaseolin. Biochemistry 23 4225-4232.

Murray, M.G., Peter, D.L., and Thompson, W.F. 1981; Ancient repeated sequences in the pea and Mung bean genomes and implications for genome evolution. J. Mol. Evol. 17 31-42.

Negoro, T., Momma, T., and Fukazawa, C. 1985; A cDNA clone encoding a glycinin A_{1a} subunit precursor of soybean. NAR 13 6719-6731.

Nevers, P., Shepherd, N.S., and Saedler, H. 1986; Plant transposable elements. Adv. Bot. Res. 12 103-203.

Nielsen, N.C. 1984; The chemistry of legumin storage proteins. Phil. Trans. R. Soc. London B. 304 287-296.

Nishioka, Y., Leder, A., and Leder, P. 1980; Unusual α-globin like gene that has clearly lost both globin intervening sequences. P.N.A.S. 77 2806-2809.

Odell, J.T., Nagy, F., and Chua, N-H. 1985; Identification of DNA sequences required for activity of the cauliflower mosaic virus 35S promoter. Nature 313 810-812.

Orgel, L.E., and Crick, F.H.C. 1980; Selfish DNA; the ultimate parasite. Nature 284 604-607.

Osborne, T.B. 1924; The vegetable proteins. Longmans Green, London.

Osborne, T.F., Gaynor, R.B., and Berk, A.J. 1982; The TATA homology and the messenger RNA 5' untranslated sequence are not required for expression of essential ElA functions. Cell 29 139-148.

Payne, P.I., and Rhodes, A.P. 1982; In: Enc. Pl. Phys. V14A 346-369. (Eds Boulter, D., and Parthier, B.), Springer Verlag Berlin Heidelburg & New York.

Peacock, D., and Boulter, D. 1975; Use of amino acid sequence data in phylogeny and evaluation of methods using computer simulation. J. Mol. Biol. 95 513-527.

Perler, F., and Efstratiadis, A. 1980; The evolution of genes : the chicken preproinsulin gene. Cell 20 555-566.

Pernollet, J-C. 1978; Protein bodies of seeds : ultrastructure, biochemistry, biosynthesis and degradation. Phytochemistry 17 1473-1480.

Pernollet, J-C., and Mosse, J. 1983; In: Seed proteins (Eds Daussant, J., Mosse, J., and Vaughan, J.), 155-191, Academic Press, New York.

Pikaard, C.S., Mignery, G.A., Ma, D.P., Stark, V.J., and Park, W.D. 1986; Sequence of two apparent pseudogenes of the major potato tuber protein, patatin. N.A.R. 14 5564-5566.

Polhill, R.M., 1981; In: Advances in legume systematics (Eds Polhill, R.M., and Raven, R.H.), Royal Botanic Gardens, Edinburgh, Scotland, 191-208.

Pribnow, D. 1975a; Bacteriophage T7 early promoters : nucleotide sequences of two RNA polymerase binding sites. J. Mol. Biol. 99 419-443.

Pribnow, D. 1975b; Nucleotide sequence of an RNA polymerase binding site at an early T7 promoter. P.N.A.S. 72 784-789.

Proudfoot, N.J. 1979; Eukaryotic promoters? Nature 279 376.

Proudfoot, N.J. 1982; The end of the message. Nature 298 516-517.

Proudfoot, N.J., and Brownlee, G.G. 1976; 3' non-coding region sequences in eukaryotic messenger RNA. Nature 263 211-214.

Proudfoot, N.J., and Maniatis, T. 1980; The structure of a human α -globin pseudogene and its relationship to α -globin gene duplication. Cell 21 537-544.

Pusztai, A., and Stewart, J.C. 1980; Molecular size, subunit structure and microheterogeneity of glycoprotein II from the seeds of kidney bean (Phaseorus vulgaris L.). Bioch. Biophys. Acta. 623 418-428.

Reddy, R., 1985; Compilation of small RNA sequences. N.A.R 13 r155-r163.

Reddy, R., and Busch, H. In: Progress in nucleic acid research and molecular biology,(Ed Cohn, W.E., Moldave, K.), Academic Press, New York, 30 127-162.

Reddy, R., Ro-Choi, T.S., Henning, D., and Busch, H. 1974; Primary sequence of U1 nuclear ribonucleic acid of Novikoff hepatoma ascites cells. J. Biol. Chem. 249 6486-6494.

Reed, R., and Maniatis, T., 1985; Intron sequences involved in lariat formation during pre-mRNA splicing. Cell 41 95-105.

Rees, H., and Jenkins, G. 1982; Assays of the phenotypic effects of changes in DNA amounts. In: Genome Evolution, (Eds Dover, G.A., and Flavell, R.B.), Academic Press, New York.

Rimpau, J., Smith, D.B., and Flavell, R.B. 1978. Sequence organisation analysis of the wheat and rye genomes by interspecies DNA/DNA hybridisation. J. Mol. Biol. 123 327-359.

Rimpau, J., Smith, D.B., and Flavell, R.B. 1980; Sequence organisation in barley and oat chromosomes revealed by interspecies DNA/DNA hybridisation. Heredity 44 131-149.

Robinson, M., Lilley, R., Little, S., Emtage, J.S., Yarranton, G., Stephens, P., Millican, A., Eaton, M., and Humphreys, G. 1984; Codon usage can affect efficiency of translation of genes in E. Coli. N.A.R. 12 6663-6671.

Rochester, D.E., Winer, J.A., and Shah, D.M. 1986; The structure and expression of maize genes encoding the major heat shock protein, hsp 70. EMBO J. 5 451-458.

Rosahl, S., Schmidt, R., Schell, J., and Willmitzer, L. 1986; Isolation and characterisation of a gene from Solanum tuberosum encoding patatin, the major storage protein of potato tubers. M.G.G. 203 214-220.

Rothman, F., Shatkin, A.J., and Perry, R.P. 1974; Sequences containing methylated nucleotides at the 5' termini of messenger RNA : Possible implications for processing. Cell 3 197-199.

Ruskin, B., Krainer, A.R., Maniatis, T., and Green, M.R. 1984; Excision of an intact intron as a novel lariat structure during pre-mRNA splicing in vitro. Cell 38 317-331.

Sanger, F., Nicklen, S., and Coulson, A.R. 1977; DNA sequencing with chain-terminating inhibitors. P.N.A.S. 74 5463-5467.

Sargan, D.R., Gregory, S.P., and Butterworth, P.H.W. 1982; A possible novel interaction between the 3'-end of 18S ribosomal RNA and the 5'-leader sequence of many eukaryotic messenger RNAs. FEBS Letts; 147 133-136.

Sasavage, N.L., Smith, M., Gillam, S., Woychick, R.P., and Rothman, F.M. 1982; Variation in the polyadenylation site of bovine prolactin mRNA. P.N.A.S. 79 223-227.

Scallon, B., Thanh, V-H., Floener, L.A., and Nielsen, N.C. 1985; Identification and characterisation of DNA clones encoding group II glycinin subunits. Theor. Appl. Genet. 70 510-519.

Schaller, H., Gray, C., Herrmann, K. 1975; Nucleotide sequence of an RNA polymerase binding site from the DNA of bacteriophage fd. P.N.A.S. 72 737-41.

Schuler, M.A., Schmitt, E.S., and Beachy, R.N. 1982; Closely related families of genes code for the α and α^1 subunits of the soybean 7S storage protein complex. N.A.R. 10 8225-8240.

Schuler, M.A., Ladin, B.F., Pollaco, J.C., Freyer, G., and Beachy, R.N. 1982; Structural sequences are conserved in the genes coding for the α , α^1 and β subunits of the soybean 7S storage protein complex. N.A.R. 10 8245-8261.

Sengupta, C., Deluca, V., Bailey, D.S., and Verma, D.P.S. 1981; Post-translational processing of 7S and 11S components of soybean (Glycine max cultivar Prize) storage proteins. Pl. Mol. Biol. 1 19-34.

Setzer, D.R., McGrogan, M., Nunberg, J.H., and Schimke, R.T. 1980; Size heterogeneity in the 3' end of dihydrofolate reductase messenger RNA in mouse cells. Cell 22 361-370.

Shah, D.M., Hightower, R.C., and Meagher, R.B. 1983; Genes encoding Actin in higher plants : intron positions are highly conserved, but the coding sequences are not. J. Mol. App. Genet. 2 111-126.

Shah, D.M., Hightower, R.C., and Meagher, R.B. 1982; Complete nucleotide sequence of a soybean Actin gene. P.N.A.S. 79 1022-1026.

Sharp, P.A. 1987; Splicing of messenger RNA precursors. Science 235 766-771.

Sharp, P.M., and Li, W-S. 1986; Codon usage in regulatory genes in Escherichia Coli does not reflect selection for rare codons. N.A.R. 14 7737-7749.

Sharp, P.M., and Li, W-H. 1987; The codon adaptation index - a measure of directional synonymous codon usage bias, and its potential applications. N.A.R. 15 1281-1295.

Sharp, P.M., Tuohy, T.M.F., and Mosurski, K.R. 1986; Codon usage in yeast : Cluster analysis clearly differentiates highly and lowly expressed genes. N.A.R. 14 5125-5143.

Shatkin, A.J. 1985; mRNA cap binding proteins : Essential factors for initiating translation. Cell 40 223-224.

Shen, S-H., Slighton, J.L., and Smithies, O. 1981; A history of the human fetal globin gene duplication. Cell 26 191-203.

Shewry, P.R., Miflin, B., and Kasarda, D.D. 1984; The structural and evolutionary relationships of the prolamin storage proteins of barley, rye, and wheat. Phil. Trans. R. Soc. Lond. B. 304 297-308.

Simon, A.E., Tenbarge, K.M., Scofield, S.R., Finkelstein, R.R., and Crouch, M.L. 1985; Nucleotide sequence of a cDNA clone of Brassica napus 12S storage protein shows homology with legumin from Pisum sativum L. Pl. Mol. Biol. 5 191-201.

Simpson, J., Van Montagu, M., and Herrera-Estrella, L. 1986; Photosynthesis associated gene families : Differences in response to tissue specific and environmental factors. Science 233 34-38.

Slightom, J.L., Blechl, A.E., and Smithies, O. 1980; Human fetal $^{G}\gamma$ and $^{A}\gamma$ globin genes : Complete nucleotide sequences suggest that DNA can be exchanged between these duplicated genes. Cell 21 627-638.

Slightom, J.L., Drong, R.F., Klassy, R.C., and Hoffman, L.M. 1985; Nucleotide sequences from phaseolin cDNA clones : the major storage proteins from Phaseolus vulgaris are encoded by two unique gene families. NAR 13 6483-6498.

Slightom, J.L., Sun, S.M., and Hall, T.C. 1983; Complete nucleotide sequence of a french bean storage potein gene : Phaseolin. P.N.A.S. 80 1897-1901.

Smithies, O., Engells, W.R., Devereux, J.R., Slightom, J.L., and Shen, S-H. 1981; Base substitutions, length differences and DNA strand assymetries in the human $^{G}\gamma$ and $^{A}\gamma$ fetal globin region. Cell 26 345-353.

Sneath, P.H.A. 1957; The application of computers to taxonomy. J. Gen. Microbiol. 17 201-226.

Sneath, P.H.A., and Sokal, R.R. 1973; Numerical taxonomy. Freeman and Co., San Francisco.

Sommer, H., and Saedler, M. 1986; Structure of the chalcone synthase gene from Antirrhinum majus. M.G.G. 202 429-434.

Southern, E.M. 1975; Detection of specific sequences among DNA fragments separated by gel electrophoresis. J. Mol. Biol. 98 503-517.

Spena, A., Viotti, A., and Pirrotta, V. 1983; Two adjacent genomic zein sequences : structure, organisation and tissue-specific restriction pattern. J. Mol. Biol. 169 799-811.

Spencer, D., Chandler, P.M., Higgins, T.J.V., Inglis, A.S., and Rubira, M. 1984; Sequence interrelationships of the subunits of vicilin from pea (Pisum sativum) seeds. Pl. Mol. Biol. 2 259-268.

Staswick, P.E., Hermodson, M.A., and Nielsen, N.C. 1981; Identification of the acidic and basic subunit complexes of glycinin. J. Biol. Chem. 256 8752-8755.

Stone, E.M., Rothblum, K.N., and Schwartz, R.J. 1985; Intron-dependent evolution of chicken glyceraldehyde phosphate dehydrogenase gene. Nature 313 498-499.

Sudhof, T.C., Goldstein, J.L., Brown, M.S., and Russell, D.W. 1985; The LDL receptor gene : A mosaic of exons shared with different proteins. Science 228 815-822.

Sun, S.M., Slightom, J.L., and Hall, T.C. 1981; Intervening sequences in a plant gene - comparison of the partial sequence of cDNA and genomic DNA of french bean Phaseolin. Nature 289 37-41.

Talbot, D.R., Adang, M.J., Slightom, J.L., and Hall, T.C. 1984; Size and organisation of a multigene family encoding Phaseolin, the major seed storage protein of Phaseolus vulgaris. M.G.G. 198 42-49.

Tautz, D., and Renz, M. 1983; An optimised freeze-squeeze method for the recovery of DNA fragments from Agarose gels. Anal. Biochem. 132 14-19.

Taylor, R.H., and Dubin, D.T. 1975; The methylation status of mammalian mitochondrial messenger RNA. J. Cell Biol. 67 428A.

Thanh, V.H., and Shibasaki, K. 1977; Beta conglycinin from soybean proteins - isolation, immunological and physicochemical properties of the monomeric forms. Bioch. Biophys. Acta. 490 370-384.

Thimmappaya, B., andSchenk, T.1979; Nucleotide sequence analysis of variable deletion mutants lacking segments of the simian virus 40 genome coding for small t antigen. J. Virol. 30 668-673.

Thomas, P.S. 1980; Hybridisation of denatured RNA and small DNA fragments transferred to nitrocellulose. P.N.A.S. 77 5201-5205.

Thompson, W.F., and Murray, M.G. 1980; Sequence organisation in pea and mung bean DNA and a model for genome evolution. Fourth John Innes Symposium, 31-45.

Timko, M.P., Kausch, A.P, Castresana, C., Fassler, J., Herrera-Estrella, L., Van den Broeck, G., Van Montagu, M., Schell, J., and Cashmore, A.R. 1985; Light regulation of plant gene expression by an upstream enhancer-like element. Nature 318 579-582.

Tischer, E., Das Sarma, S., and Goodman, H.M. 1986; Nucleotide sequence of an alfalfa glutamine synthetase gene. M.G.G. 203 221-229.

Tollervey, D., and Mattaj, I.W. 1987; Fungal small nuclear ribonucleoproteins share properties with plants and verebrate U-snRNPs. EMBO J. 6 469-476.

Tosi, M., Young, R.A., Hagenbuchle, O., and Schibler, V. 1981; Multiple polyadenylation sites in a mouse α -amylase gene. N.A.R. 9 2313-2323.

Tumer, N.E., Thanh, V.H., and Nielsen, N.C. 1981; Purification and characterisation of mRNA from soybean seeds. J. Biol. Chem. 256 8756-60.

Vanin, E.F. 1984; Processed pseudogenes - characteristics and evolution. Bioch. Biophys. Acta. 782 231-241.

Vanin, E.F., Goldberg, G.I. Tucker, P.W., and Smithies, O. 1980; A mouse α -globin related pseudogene lacking intervening sequences. Nature 268 222-226.

Vieira, J., and Messing, J. 1982; The pUC plasmids, an M13mp7 - derived system for insertion mutagenesis and sequencing with synthetic universal primers. Gene 19 259-268.

Vijayraghavan, U., Parker, R., Tamm, J., Iimura, Y., Rossi, J., Abelson, J., and Guthrie, C. 1986; Mutations in conserved intron sequences affect multiple steps in yeast splicing pathway, particularly assembly of the spliceosome. EMBO J. 5 1683-1695.

Vodkin, L.O., Rhodes, P.R., and Goldberg, R.B. 1983; cA lectin gene insertion has the structural features of a transposable element. Cell 34 1023-1031.

Von Heijne, G. 1985; Signal sequences - the limits of variation. J. Mol. Biol. 184 99-105.

Voss, S.D., Schlokat, U., and Gruss, P. 1986; The role of enhancers in the regulation of cell-type specific transcriptional control. Trends in Bioch. Sci. 11 287-289.

Wahli, W., Dawid, I.B., Ryffel, G.U., and Weber, R. 1980; Comparative analysis of the structural organisation of two closely related vitellogenin genes in Xenopus laevis. Cell 20 107-117.

Wallace, J.C., and Edmonds, M. 1983; Polyadenylated nuclear RNA contains branches. P.N.A.S. 80 950-954.

Waterhouse, R.N. 1985; Ph.D. Thesis, University of Durham.

Weber, E., and Neumann, D. 1980; Protein bodies - storage organelles in plant seeds. Biochem. Physiol. Pflanz. 175 279-306.

Weidemann, L.M., and Perry, R.P. 1984; Characterisation of the expressed gene and several processed pseudogenes for the mouse ribosomal protein L30 gene family. Mol. Cell. Biol. 4 2518-2528.

Weiher, H., Konig, M., and Gruss, P. 1983; Multiple point mutations affecting the simian virus 40 enhancer. Science 219 626-631.

Weiss, B., Jacquemin-Sablon, A., Live, T.R., Fareed, G.C., and Richardson, C.C. 1968; Enzymatic breakage and joining of deoxyribonucleic acid.VI.Further purification and properties of polynucleotide ligase from Escherichia Coli infected with Bacteriophage T_4. J. Biol. Chem. 243 4543-4555.

Werr, W., Frommer, W-B., Maas, C., and Starlinger, P. 1985; Structure of the sucrose synthetase gene on chromosome 9 of Zea mays L. EMBO J 4 1373-1380.

Wiborg, O., Hyldig-Nielsen, J.J., Jensen, E.O., Paludan, K., and Marcker, K.A. 1982; The nucleotide sequences of two leghaemoglobin genes from soybean. N.A.R. 10 3487-3493.

Wickens, M., and Stephenson, P. 1984; Role of conserved AAUAAA sequence - four AAUAAA point mutations prevent mRNA 3' end formation. Science 226 1045-1051.

Wienand, U., Langridge, P., and Feix, G. 1981; Isolation and characterisation of a genomic sequence of maize coding for a zein gene. M.G.G. 182 440-444.

Wobus, U., Baumlein, H., Bassuner, R., Grafe, R., Jung, R., Muntz, K., Saalbach, G., and Weschke, W. 1986; Cloning and characterising *Vicia faba* seed storage protein genes. Biol. Zbl. 105 121-128.

Yanisch-Peron, C., Vieira, J., and Messing, J. 1985; Improved M13 phage cloning vectors and host strains : nucleotide sequences of the M13 mp18 and pUC 19 vectors. Gene 33 103-119.

Young, R.A., Hagenbuchle, O., and Schibler, U. 1981; A single mouse α-amylase gene specifies two different tissue specific mRNAs. Cell 23 451-458.

Zeitlin, S., and Efstratiadis, A. 1984; *In vivo* splicing products of the rabbit β-globin pre-mRNA. Cell 39 589-602.

Zhuang, Y., and Weiner, A.M. 1986; A compensatory base change in Ul snRNA suppresses a 5' splice site mutation. Cell 46 827-835.

Zuckerlandl, E., and Pauling, L. 1965; Evolving genes and proteins. (Eds. Bryson, V., and Vogel, H.J.), Academic Press, New York, 97-166.

* Boulter D. 1984; Cloning of pea storage protein genes.Phil.Trans.R.Soc.Lond.B.304 323-332.

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