Studies on the regulation of expression
of the
7S storage proteins from the garden pea,
_Pisum sativum._

by
Edward James Newbiggin

A thesis submitted for the degree of Doctor of Philosophy.

Department of Botany,
The Faculties,
Australian National University.
Statement

I declare that the results presented in this thesis are my own work except where otherwise stated and were obtained under the supervision of Dr T.J. Higgins. None of the material has been presented for the award of any other diploma or degree in any other university.

I willingly acknowledge the contributions of the following workers to chapter 3 which has been prepared as a paper for publication in Planta:

Dr B. deLumen:- Isolation and restriction mapping of the pea convicilin gene; construction of the binary vector, BDL5.
Dr P. Chandler:- Sequencing of the convicilin cDNA, pPS15.28.
Dr A. Gould:- Construction of the binary vector, pAG1.
Mr J. March:- Isolation of convicilin protein from pea seeds.
Drs R. Blagrove and A. Kortt:- N-terminal sequencing of convicilin and determinations of oligomeric structure.

Other contributions of experimental materials are recorded at the appropriate place in the text.

Some of the data collected during the period of my Ph.d. was published as a paper (Plant Molec.Biol 11, 683-695, 1988) and is presented here as appendix 1. My contributions to this paper included isolation of the pEN2 (vicilin) gene and characterization of the gene by restriction mapping and DNA sequencing, preparation of RNA from transgenic tobacco seeds and S1 mapping of the transcripts of this gene in pea and tobacco.

Edward Newbigin

This thesis is dedicated to my family, friends and B3.

...I saw the pale student of unhallowed arts kneeling beside the thing he had put together.

The diary of Mary Wollstonecraft Shelly.
Acknowledgements

I wish to thank the three people who acted, at various stages of my study, as my supervisors, TJ, Peter and Malcolm, especially TJ. Without his invaluable and friendly help, this thesis would not have occurred. I also wish to note the great contribution made by the folk in the lab, both those currently there and those who have moved on, not only to my research but also to my mental well-being. The warmth and humour of your companionship has been one of the perquisites of my student days (along with the cheap travel and good discounts to films, shows and bands). To Uncle Don, you were always one to offer wise and well-considered advice when it was most needed. To the community of happy souls at Plant Industry, whose liveliness and infectious enthusiasm for science has left a deep impression on me. To Sarah, for all the years of nest-building and those ephemeral house mates who passed through Jarrah St at some time adding exotic touches as they went. To my cheer leader, Lindy. To the people with whom I played soccer or tennis, went eating, drinking or bush walking, talked to in English or Chinese, to anybody who helped me or I helped, I extend a heart-felt "thank you" for making the last few years so pleasant.

I extend my appreciation to Dr Peacock for making the facilities at Plant Industry available to me and to Emile, Dave and Sandie for their help with the photography and illustrations.

I gratefully acknowledge the receipt of a Commonwealth Postgraduate Scholarship.
Abstract

The pea seed storage protein, vicilin, is synthesized as two size classes of primary translation products. Some members of the class that are synthesized as Mr 50,000 polypeptides (also called the vicilins) are subsequently cleaved to a number of smaller molecules whereas no such post-translational processing of the Mr 70,000 polypeptide (called convicilin) appears to occur.

Two genes, one encoding a representative of the Mr 50,000 vicilins and the other convicilin, were isolated from a pea genomic library and characterized by restriction mapping and DNA sequencing. Both genes were subsequently integrated into the genome of tobacco by Agrobacterium tumefaciens-mediated transformation and the fidelity of the expression of the genes in the new host species was examined. Both the genes were functional in tobacco although the convicilin gene was expressed to a greater extent than the vicilin gene. The two genes were found to be expressed only in tobacco seeds and the temporal features of this expression were similar to that found in pea. Transcription of these genes was initiated at similar positions in both pea and tobacco.

In tobacco seeds, the protein products of the two pea genes were processed to lower molecular weight polypeptides, a feature characteristic of pea seed vicilin but not convicilin.

Although the coding regions of the two pea storage protein genes showed a high level of similarity, not only to each other but also to vicilin-like genes from other legumes, the 5' flanking sequences of the genes were largely divergent. Sequence conservation in the upstream region was restricted to a stretch of thirty nucleotides that were common to the two pea genes and the vicilin-like genes from soya bean (α', β-conglycinin) and french bean (β-phaseolin). This region of conserved sequence (called the vicilin box) corresponded to a region upstream of the α', β-conglycinin gene that was able to correctly express this gene in transgenic plants.
To determine whether the vicilin box acted as a seed promoter, a fragment of the upstream region of the vicilin gene containing this sequence was ligated to a reporter gene (Chloramphenicol Acetyl Transferase) and introduced into the tobacco genome by Agrobacterium-mediated transformation. The seeds of transgenic plants were assayed for CAT activity and the results compared with those obtained from plants transformed with a similar construction carrying a larger fragment of the vicilin upstream sequence. Plants that had integrated the former construction into their genomes did not have detectable levels of CAT in their seeds whereas CAT was detectable in the seeds of tobacco plants transformed with the latter construction. This result demonstrated that the vicilin box was insufficient for seed expression and that sequences further upstream were required. A model of the organization of the vicilin promoter is proposed and experiments designed to test this model are discussed.
# Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Adenine nucleotide</td>
</tr>
<tr>
<td>C</td>
<td>Cytosine nucleotide</td>
</tr>
<tr>
<td>G</td>
<td>Guanine nucleotide</td>
</tr>
<tr>
<td>T</td>
<td>Thymine nucleotide</td>
</tr>
<tr>
<td>ABA</td>
<td>abscisic acid</td>
</tr>
<tr>
<td>NAA</td>
<td>naphthalene acetic acid</td>
</tr>
<tr>
<td>BA</td>
<td>benzyladenine</td>
</tr>
<tr>
<td>SDS</td>
<td>sodium dodecyl sulphate</td>
</tr>
<tr>
<td>CAM</td>
<td>chloramphenicol</td>
</tr>
<tr>
<td>Ac</td>
<td>acetylated derivatives of chloramphenicol</td>
</tr>
<tr>
<td>EDTA</td>
<td>ethylenediaminetetra-acetic acid</td>
</tr>
<tr>
<td>Tris</td>
<td>tris(hydroxymethyl)aminomethane</td>
</tr>
<tr>
<td>PTH</td>
<td>phenylthiohydantoin</td>
</tr>
<tr>
<td>DTSP</td>
<td>dithiobis(succinimidyl propionate)</td>
</tr>
<tr>
<td>SSC</td>
<td>standard saline citrate</td>
</tr>
<tr>
<td>bp</td>
<td>base pair</td>
</tr>
<tr>
<td>kb</td>
<td>kilobase pairs</td>
</tr>
<tr>
<td>Mr</td>
<td>relative molecular weight</td>
</tr>
<tr>
<td>DAF</td>
<td>days after flowering</td>
</tr>
<tr>
<td>ER</td>
<td>endoplasmic reticulum</td>
</tr>
<tr>
<td>SDS-PAGE</td>
<td>SDS polyacrylamide gel electrophoresis</td>
</tr>
<tr>
<td>cat</td>
<td>chloramphenicol acetyl transferase gene</td>
</tr>
<tr>
<td>CAT</td>
<td>chloramphenicol acetyl transferase enzyme</td>
</tr>
<tr>
<td>nos</td>
<td>nopaline synthase gene</td>
</tr>
<tr>
<td>mRNA</td>
<td>messenger RNA</td>
</tr>
<tr>
<td>tRNA</td>
<td>transfer RNA</td>
</tr>
<tr>
<td>cDNA</td>
<td>copy DNA</td>
</tr>
<tr>
<td>HPLC</td>
<td>high pressure liquid chromatography</td>
</tr>
<tr>
<td>CaMV</td>
<td>cauliflower mosaic virus</td>
</tr>
</tbody>
</table>
## Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Statement</td>
<td>(ii)</td>
</tr>
<tr>
<td>Acknowledgements</td>
<td>(iv)</td>
</tr>
<tr>
<td>Abstract</td>
<td>(v)</td>
</tr>
<tr>
<td>Abbreviations</td>
<td>(vii)</td>
</tr>
</tbody>
</table>

**Chapter 1: Introduction.**

<table>
<thead>
<tr>
<th>Subsection</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>General considerations</td>
<td>2</td>
</tr>
<tr>
<td>Structure and homologies between 7S storage proteins</td>
<td>3</td>
</tr>
<tr>
<td>Biosynthesis and transport of 7S storage proteins</td>
<td>9</td>
</tr>
<tr>
<td>Regulation of 7S storage proteins by mineral nutrients and phytohormones</td>
<td>12</td>
</tr>
<tr>
<td>7S globulin gene structure and chromosomal location</td>
<td>15</td>
</tr>
<tr>
<td>Thesis outline</td>
<td>18</td>
</tr>
</tbody>
</table>

**Chapter 2: Mapping of the transcription start-points of two pea seed storage protein genes and the identification of potential regulatory sequences in the 5' sequence.**

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction</td>
<td>21</td>
</tr>
<tr>
<td>Methods</td>
<td></td>
</tr>
<tr>
<td>DNA sequencing and sequence comparisons</td>
<td>22</td>
</tr>
<tr>
<td>S1 mapping of vicilin and convicilin transcripts</td>
<td>22</td>
</tr>
<tr>
<td>Results</td>
<td>24</td>
</tr>
<tr>
<td>Discussion</td>
<td>27</td>
</tr>
</tbody>
</table>

**Chapter 3: Pea convicilin: Structure and primary sequence of the protein and its expression in the seeds of transgenic tobacco.**

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction</td>
<td>31</td>
</tr>
</tbody>
</table>
Methods

Isolation of a convicilin gene and sequencing of cDNA and genomic fragments 33
Purification, oligomeric structure and N-terminal sequencing of convicilin 33
Transfer of convicilin gene to tobacco 34
Protein detection by Western blot 34
DNA isolation and Southern blotting of genomic DNA 34
RNA isolation, S1 mapping and densitometry 35

Results

Isolation and structure of convicilin 35
Convicilin sequence and structure comparisons with vicilin 37
Expression of pea convicilin in tobacco 39
Developmental regulation and organ specificity 41

Discussion 42

Chapter 4: Expression of vicilin promoter:cat chimeric genes in the seeds of transgenic tobacco plants.

Introduction 47

Methods

Construction of vicilin promoter:cat chimeric genes 48
Transformation of tobacco plants 49
Assays for CAT activity in transgenic tobacco seeds 49
DNA isolation 49

Results

Assay of transgenic plants for presence of chimeric genes 50
Expression of vicilin promoter:cat chimeric gene constructions in tobacco 52

Discussion 54
Appendix 1: The sequence of a pea vicilin gene and its expression in transgenic tobacco plants.

Abstract
Introduction

Methods

- Construction and screening of a pea genomic library and sequencing of a vicilin gene
- Transfer of the pea vicilin gene into tobacco
- Isolation and characterization of nucleic acids from tobacco
- Western blotting
- Densitometry
- Immunocytochemistry

Results

- The structure of a pea vicilin gene
- The expression of a pea vicilin gene in tobacco
- Vicilin gene copy number
- Intracellular localization of pea vicilin in tobacco seeds
- Organ specificity of vicilin gene expression
- Temporal pattern of vicilin expression in developing transgenic tobacco seeds
- The fate of pea vicilin in transgenic tobacco seeds during germination

Discussion

Acknowledgements

References
Chapter 1

Introduction
General considerations

The term "seed storage protein" is used to describe those proteins laid down in the seed whose function is to provide, upon hydrolysis, fixed carbon and nitrogen to sustain the early heterotrophic growth of the seedling. Osborne (1924) in his pioneering work on the characterization of seed proteins, recognized that these proteins, regardless of the species from which they were obtained, could be divided into four solubility classes, namely albumins, globulins, prolamine and glutelins. Proteins of the albumin class were soluble in water whereas the globulins were soluble in salt solutions (0.5 to 1M NaCl). Prolamines could be extracted from the finely ground seed meal with aqueous alcohol (50 to 70%) and the glutelins with either acidic or alkaline solutions. After examining the solubility of seed proteins from many species, Osborne concluded that, in general, the major proteins of legumes were globulins while those of cereals were glutelins and prolamine. More recently, exceptions to this generalization have been found, notably in oats where globulins are the major storage class (Peterson 1978).

The albumin fraction of seeds is not typically thought to play a major role as a storage reserve. This fraction contains enzymes and other proteins essential for the maintenance of normal cell metabolism and function. However, some of these proteins are present in sufficiently large amounts to suggest that in some species they may have a secondary role as storage proteins. Examples of this may be urease and concanavalin A from jack bean (Bailey and Boulter 1971; Hague 1975), lectin from french bean (Bollini and Chrispeels 1978) and β-amylase and the chymotrypsin inhibitors in some genotypes of barley (Boisen et al. 1981). Moreover, there are legumes, such as castor bean and pea, where significant amounts of small, water-soluble proteins are found (Youle and Huang 1978; Schroeder 1984 a) that appear to function only as storage proteins (Higgins et al. 1986).

The amino acid compositions of the major storage proteins of cereals are characterized by high amounts of Glx (glutamine and glutamate) and proline but are very low in basic amino acids, especially lysine. The legume storage proteins are also rich in Glx and have relatively
high proportions of arginine, lysine and Asx (asparagine and aspartate) and low amounts of sulphur-containing amino acids (cysteine and methionine). This predominance of nitrogen-rich amino acids presumably reflects the functional role of these proteins as reserves to be used during seedling development.

At the end of the last century, Osborne and Campbell (1896) reported that there were two distinct kinds of globulins in the seeds of leguminous plants. The two fractions were named legumin and vicilin. This early observations were extended by Danielsson (1949) and the purification and taxonomic distribution of the different globulins in legumes was reviewed later by Derbyshire et al. (1976). The two kinds of globulin were distinguished as discrete size classes with sedimentation coefficients (S values) of 7 to 8S (vicilin-type) and 11 to 14S (legumin-type) (Danielsson 1949). Each of these size classes consists of large, oligomeric proteins made up of closely-related polypeptides. The subunits of an oligomer are derived from one of two families of globulin genes. The members of one of these families, the vicilin family, are the topic of this thesis and I should like to limit my discussion to them from this point. However, many of the statements made about this family may be equally applicable to other classes of storage proteins.

Most of the information available on the 7S globulins comes from studies on pea (Pisum sativum), french bean (Phaseolus vulgaris) and soya bean (Glycine max) and so this review will largely be limited to them although other species may be occasionally mentioned. The trivial names of the 7S globulins from the three above-mentioned species are vicilin, phaseolin and β- and γ-conglycinin, respectively. Some confusion may arise from the use of the Greek alphabet to describe either the native protein (e.g. β-, γ-conglycinin), polypeptide subunits of the native protein (e.g. α-, β- and γ- subunits of phaseolin; α-, α'- and β- subunits of β-conglycinin) or subfragments of a polypeptide (e.g. the α, β and γ polypeptides of vicilin). These terms will be used with care and it is hoped that any ambiguity will be minimal.

**Structures of and homologies between 7S storage proteins**

Although characterized as having a sedimentation coefficient of 7S, these proteins can
associate/dissociate to give proteins of 3S, 7S, 9S or 18S with variation of either ionic strength or pH (Derbyshire et al. 1976). The quaternary structure is based on a trimer of polypeptide subunits (Gatehouse et al. 1981; Thanh and Shibasaki 1977), or multiples of a trimer such as the case of the 18S form of phaseolin (Sun et al. 1974). There are two possible exceptions to this. One is convicilin from pea (a component of the vicilin fraction) which is reported to be a tetramer based on its apparent molecular weight as determined by gel-filtration chromatography (Croy et al. 1980) although other reports show convicilin to be a trimer (Casey and Sanger 1980; chapter 3 of this thesis). The other is a basic 7S globulin from soya bean which is composed of four pairs of Mr 26,000 and Mr 16,000 subunits and accounts for about 3% of the total seed protein (Hirano et al. 1987; Kagawa et al. 1987). The subunits, which are linked together by disulphide bridge(s) (an unusual feature for 7S globulins), have been found in other legume species (Kagawa et al. 1987).

The size of the oligomers can vary appreciably. For example, those from pea and soya bean can be resolved as a number of distinct proteins with molecular weights between Mr 150,000 and 190,000 (Thomson et al. 1980; Thanh and Shibasaki 1978). The number and size of these proteins varies within the species and genetic evidence indicates that these variants were genotypë-specific and inherited in a co-dominant manner (Thomson and Schroeder 1978).

The heterogeneity in the size of the oligomers is a reflection of the heterogeneity in the size of the subunits. This variation in the subunits results from one or a combination of the following factors: sequence differences among the constituent polypeptides, presence or absence of covalently-bound carbohydrate and proteolytic processing of a precursor. Different forms of the oligomers result from various combinations of the heterogeneous subunits and attempts have been made to define the polypeptide composition of some of the 7S aggregates (Thomson et al. 1980; Thanh and Shibasaki 1976, 1978). At least in the case of pea, the distinct subunit make-up of each of the different vicilin oligomers is genetically determined (Thomson et al. 1980).
As mentioned above, subunit diversity has a number of causes. However, these causes do not apply equally in all species as the degree of glycosylation and sequence heterogeneity are sufficient to account for most of the subunits in French bean and soya bean, whereas proteolytic processing is more important in pea and *Vicia faba* (broad bean).

Phaseolin, which accounts for 50% of the total protein from French beans at maturity (Bollini and Chrispeels 1978), is resolved into three classes of polypeptide by SDS polyacrylamide gel electrophoresis (SDS-PAGE). These classes are referred to as α, β, and γ and have molecular weights of Mr 51,000, 48,000 and 45,000, respectively (Brown *et al.* 1981 a). However, phaseolin mRNA isolated from bean cotyledons directs the synthesis of only non-glycosylated α and β subunits in a cell-free, wheat germ translation system (Hall *et al.* 1978). The γ subunit seems to be a result of partial glycosylation of a β-type polypeptide (Bollini *et al.* 1983; Slightom *et al.* 1985) and consistent with this, it has a smaller carbohydrate to protein ratio than either the α or β subunits (Hall *et al.* 1977).

The sequences of several cDNA clones for phaseolin are known and these have been classified as encoding either α or β polypeptides (Slightom *et al.* 1985). The sequences show that the larger α polypeptides differ from the β-type mainly by the inclusion of two small repeated sequences, adding a total of 14 amino acids. Apart from this difference, the α- and β-type phaseolin cDNAs are 98% identical, consistent with their derivation from a common ancestral gene. Subunit heterogeneity may be further increased by proteolytic processing as a number of different N-termini have been reported for the mature protein (Slightom *et al.* 1985).

There are three types of native storage protein from soya bean that have been termed conglycinin (Catsimpoolas and Ekenstam 1969). One of these, α-conglycinin, is a 2S protein that has received scant attention since its first description. The remaining conglycinins (β- and γ-type) are both 7S proteins with γ-conglycinin accounting for only 10% of the total 7S fraction. The major 7S globulin, β-conglycinin, is made up of various combinations of three types of subunits (see previous comments) which are called α (Mr 76,000), α' (Mr 83,000) and β (Mr 53,000) (Thanh and Shibasaki 1976; Beachy *et al.* 1981). Other minor subunits also have
been reported (Thanh and Shibasaki 1976; Coates et al. 1985). All of the β-conglycinin subunits are glycoproteins with the α' and α polypeptides containing about twice the carbohydrate of the β polypeptides (Thanh and Shibasaki 1976, 1977). Two-dimensional gels resolve the α, α' and β subunits into several species that differ in charge but not in molecular weight (Ladin et al. 1984, 1987). Analysis of soya bean cultivars that do not accumulate the α' subunits (Ladin et al. 1984) and expression of isolated α' and β subunit genes in transgenic tobacco (Beachy et al. 1985; Bray et al. 1987) indicated that charge heterogeneity can arise from a single polypeptide and need not be caused by a number of closely-related polypeptides.

Amino acid sequence comparisons of β-conglycinin α' and β subunits showed a high degree of similarity between these two proteins with the exception of 179 residues near the amino terminus of the α' subunit that were not found in the β (Doyle et al. 1986; Tierney et al. 1987). This large insertion accounts for most of the difference in molecular weight observed between the α' and β polypeptides. When this insert was not considered, then the α' and β sequences were approximately 75% identical.

Sequence analysis of cDNAs corresponding to the α and α' polypeptides of β-conglycinin has shown that they too are highly similar and differ at only 6% of their nucleotides (Schuler et al. 1982a). One interesting observation was that a sequence of 155 bp, common to the coding regions of α and α' cDNAs, was also in the 3' non-coding region of a cDNA clone for a glycinin (11S) precursor (Schuler et al. 1982b; Scallon et al. 1985). In their paper, Schuler et al. (1982b) speculate that this sequence may play a role in the expression, structure or stability of seed mRNAs.

Whereas most of the heterogeneity seen in the 7S globulin subunits from soya bean and french bean can be accounted for by differences in amino acid sequences or in the level of glycosylation, this is not the case with the vicilin polypeptides from pea, broad bean or mung bean (Vigna radiata) where proteolytic processing of precursors is more important (Gatehouse et al. 1983; Spencer et al. 1983; Scholz et al. 1983; Ericson and Chrispeels 1976).
Pulse labelling of pea cotyledons showed two primary translation products for vicilin of Mr 70,000 and 50,000 (Chrispeels et al. 1982 a; Gatehouse et al. 1981). The larger of these two polypeptides, called convicilin (Croy et al. 1980), does not appear to undergo any further processing (Chrispeels et al. 1982a). Some or all of the Mr 50,000 polypeptides act as precursors for the smaller subunits observed in vivo by SDS PAGE (Chrispeels et al. 1982a; Gatehouse et al. 1981). These smaller subunits have sizes of: Mr 34,000, 30,000, 25,000, 18,000, 14,000, 13,000 and 12,000 (Spencer et al. 1983). Based on protein and cDNA sequence information, a scheme for the derivation of the smaller vicilin subunits from the precursors has been proposed (Gatehouse et al. 1983; Spencer et al. 1983). The scheme suggests the presence of two potential processing sites within the Mr 50,000 polypeptide which would cleave it into molecules of Mr 18,000, 13,000 and 12,000. Gatehouse et al. (1983) have named these the α, β and γ subunits, respectively. Processing at either one of the two sites would give rise to the polypeptides of intermediate size observed in the mature seed. Previously it had been shown that a polypeptide of Mr 14,000 was a glycosylated form of the Mr 12,000 (γ) subunit (Davey et al. 1981; Badenoch-Jones et al. 1981). Only this polypeptide and some of the Mr 50,000 molecules contain carbohydrate. The proposed processing scheme does not make clear how the Mr 34,000 subunit is derived.

Lycett et al. (1983) suggested that the vicilin precursor could be divided into four groups depending on whether they underwent processing at one, both or neither of the two presumed cleavage sites. However, experiments in which a number of 7S storage protein genes (including vicilin) have been transferred to new host species have demonstrated that proteolytic cleavage need not process all the available sites in all the available precursor molecules as both processed and unprocessed polypeptides of the same gene product can exist in the seed together (Sengupta-Gopalan et al. 1985; Beachy et al. 1985; Bray et al. 1987; this thesis, chapter 3 and appendix 1).

Comparisons of protein, cDNA and genomic DNA sequences of vicilin and convicilin have shown sequence similarity between the subunits of these two 7S proteins (Casey et al. 1984; Bown et al. 1988). The major difference being that convicilin polypeptides have extra coding
sequences at their N-terminus (Casey et al. 1984; Bown et al. 1988; this thesis, chapter 3) and thus the relationship between them is analogous to that between the α/α' and β subunits of β-conglycinin.

A number of cDNAs coding for the Mr 50,000 subunits have been isolated (Lycett et al. 1983; Chandler et al. 1984). These can be divided into three families on the basis of sequence similarity with the members of a family showing between 15 and 25% nucleotide difference with members of another family (Chandler et al. 1984). Sequences within a family are less than 5% divergent.

As indicated above, sequences (protein, cDNA and genomic) of representatives of most of the major classes of 7S globulin subunit are available. Interspecific comparisons of these sequences have shown fundamental similarities that are in agreement with the idea that these genes are descended from a common ancestor (Argos et al. 1985; Doyle et al. 1986; Borroto and Dure 1987). Despite this, the proteins lack serological cross-reactivity (Klozova and Kloz 1972; Dudman and Millerd 1975; Schuler et al. 1983) and have dissimilar gel electrophoresis patterns. This, together with the observation of Schuler et al. (1983) that there is a higher degree of similarity between 7S subunits when DNA rather than protein sequences are compared suggests that selection has preserved overall protein structure but not specific amino acid sequences. There is experimental evidence to show that at least one region of conserved structure may be involved in the association of subunits to form oligomers. Argos et al. (1985) identified, on the basis of predicted secondary structure, an area of mostly hydrophobic residues that was preserved in the C-terminal regions of both 11S and 7S storage proteins. Dickinson et al. (1987) subsequently showed, using an in vitro system, that deletion of this conserved structure removed the ability of glycinin (the 11S storage protein from soya bean) subunits to assemble into oligomers. Modification of other areas of the protein did not have this effect suggesting that changes to regions other than the area of preserved structure could be tolerated with minimal effect upon the structure of the protein. The true importance of specific domains within the subunits to overall storage protein structure however will have to await crystallographic determination of the three-dimensional structure of the oligomers.
Biosynthesis and transport of 7S storage proteins

The previous section dealt mostly with the features of 7S globulin proteins as they are observed in fully-ripe seeds. This section will review the biochemical events preceding the formation of mature oligomers and place these events within the framework of development of the storage (cotyledonary) tissues.

During the course of normal seed development, the cotyledons pass through three successive phases. The first phase is one of cell division, followed by a phase of cell expansion and then seed desiccation and maturation. During the second of these stages, storage reserves (including proteins) accumulate and are deposited in discrete organelles. Storage proteins are deposited in small (1-5μm) membrane-enclosed organelles called protein bodies. The other major reserves laid down at this time, either carbohydrate or lipid (depending on the species) are stored as starch grains or as lipid bodies. Together these organelles become the dominant structural features by the end of the phase of cell expansion which marks the completion of the synthesis and deposition of storage reserves.

The appearance of storage proteins during this phase of seed development has been correlated to the appearance of specific storage protein mRNAs (Goldberg et al. 1981; Chandler et al. 1984) and the onset of transcription of storage protein genes (Beach et al. 1985; Walling et al. 1986). The synthesis of the major storage protein families in pea and soya bean is not co-ordinately regulated with some proteins (and their mRNAs) appearing earlier than others. Even within a family, such as the 7S globulins, all the subunits are not made at the same time. In pea, vicilin polypeptides appear about 5 days earlier than convicilin (Chandler et al. 1984) and in soya bean, α and α' subunits appear several days before the β-subunits (Meinke et al. 1981). Experiments with transfer of genes for these subunits to new hosts have shown that the information required for this separate temporal regulation is held within the genes coding for these proteins (Beachy et al. 1985; Bray et al. 1987; this thesis, chapter 3 and appendix 1).
Typical of proteins destined for transport, storage proteins are synthesized on membrane-bound polysomes as precursors containing a N-terminal extension called the signal sequence (Higgins and Spencer 1977; Bollini and Chrispeels 1979; Puchel et al. 1979; Sengupta et al. 1981; Hurkman and Beevers 1982). The function of the signal sequence is to facilitate the translocation of the polypeptide chain into the lumen of the endoplasmic reticulum (ER) as the first step in intracellular transport (Kreil 1981). The signal is co-translationally removed and there is good evidence to show that newly synthesized storage proteins accumulate transiently within the ER prior to transportation (Chrispeels et al. 1982 a,b). In cases where in vivo or in vitro translation experiments have not been able to demonstrate the existence of an extended polypeptide chain due to interference from other co-translational events (see below), sequencing of genomic DNA has shown the presence of a nucleotide sequence, 5' to the bases encoding the determined N-terminus of the mature protein, coding for a potential signal peptide (Doyle et al. 1986).

The addition of carbohydrate to 7S polypeptide chains is another event that occurs during or immediately after translation (Sengupta et al. 1981; Chrispeels et al. 1982 b; Bollini et al. 1983). There are three potential glycosylation sites (made up of the sequence: asparagine-most residues-threonine or serine) found in α', β-conglycinin and two in β-phaseolin (Doyle et al. 1986) although it seems likely that one of the sites in the α'-subunit is not used as it has the sequence asparagine-proline-serine which is rarely glycosylated (Mononen and Karjalainen 1984). Only some of the vicilin polypeptides are glycosylated and these contain a single site in a region corresponding to the Mr 12,000 (γ) subunit (see above; Lycett et al. 1983). It was suggested that because the two glycosylation sites in phaseolin are in regions with different hydropathic environments that one of them may be more accessible for addition of carbohydrate than the other (Slightom et al. 1985). This, in turn, may lead to polypeptide heterogeneity, such as was seen with the γ-type subunits. However, specific processing of the carbohydrate side chains occurs during transportation and deposition (Sturm et al. 1987), adding to subunit heterogeneity. Whatever the reason for some proteins being glycosylated, it is obvious that glycosylation is not required for subsequent steps (transport, oligomer formation) as both non-glycosylated subunits (such as some of the vicilins from pea and all
11S subunits) and those whose glycosylation has been prevented (Badenoch-Jones et al. 1981; Chrispeels et al. 1982 b) or modified (Chrispeels and Vitale 1985) by inhibitors were processed normally.

Formation of oligomers from the newly-synthesized 7S subunits is an early event which occurs soon after the entry of the polypeptide into the lumen of the ER (Chrispeels et al. 1982 a). The path by which the recently-assembled oligomers are transported through the cell and deposited in the protein body is not certain but immuno-localization studies (Baumgartner et al. 1980; Craig and Goodchild 1984 a) have shown the presence of 7S polypeptides in the Golgi complexes as well as in the ER and protein bodies. In plant cells, proteins that are destined for export or internal secretion into the vacuole (including protein bodies, see below) have a common pathway that passes from the ER and through the Golgi complex. From this point the pathways diverge and proteins are targeted to the appropriate location, either the cell wall or the vacuole (Gardiner and Chrispeels 1975; Chrispeels 1983). It seems likely that the Golgi apparatus also is involved in the transport of 7S globulins. This assumption is supported by evidence that the sodium ionophore, monensin, which affects transport through the Golgi complex, causes mistargeting of pea storage proteins to the plasma membrane (Craig and Goodchild 1984 b). It also has been suggested that a second signal, found on all proteins that are deposited in the vacuole or protein body, is required for correct targeting (Dorel et al. 1988). Protein transport to the lysosomes of some animal cells and to the vacuoles of yeast requires a second targeting signal (Johnson et al. 1987; Valls et al. 1987; von Figura and Hasilik 1986) and it seems probable that a similar situation exists in plants although the exact nature of the signal is yet to be determined. The mechanism by which storage proteins leave the Golgi apparatus and are transported to the vacuole is unclear although Golgi-derived vesicles have been thought of as a possible vehicle (Dieckert and Dieckert 1976; Harris 1984).

The numerous protein bodies seen by microscopy in cross-sections of mature seeds are derived from the one or two large vacuoles seen early in the cell expansion phase (Craig et al. 1980 a, b; Yoo and Chrispeels 1980). It appears that the fragmentation of the large vacuole is due to the progressive and extensive modification that it undergoes as a result of protein
deposition. This causes the vacuole to become a highly convoluted and complex structure that fragments to form the small individual protein bodies seen in the mature seed.

Once deposited in the protein body, the 7S polypeptides undergo further modifications. The most obvious of which is the proteolytic cleavage of some of the pea vicilin subunits (see previous comments). Other minor modifications such as changes to the carbohydrate side-chains or N- and C-terminal processing of some vicilin and phaseolin polypeptides also may occur at this time (Sengupta et al. 1981; Spencer et al. 1983; Lycett et al. 1983; Slightom et al. 1985; Sturm et al. 1987).

Regulation of 7S storage proteins by mineral nutrients and phytohormones.

The synthesis of storage proteins, as was mentioned previously, is restricted to a period of seed development just prior to the onset of quiescence. This section will review environmental and hormonal factors that have been shown to influence the production of storage proteins generally and 7S globulins in particular.

Storage proteins are the dominant protein fraction of mature seeds and, during the time that they are accumulated, most of the nitrogen supplied to the seed is channelled into their synthesis. It was therefore not surprising that supplying nitrogenous fertilizers during flowering and early seed development led to increased seed and protein production per area (Schilling 1977; Merbach 1983). This effect of late nitrogen dressing was not seen with all legumes as broad beans and yellow lupins (Lupinus luteus) do not respond although soya beans, peas and white lupins (Lupinus albus) do. It is not known whether there is a general increase in the synthesis of all storage proteins under these conditions or if there are certain subfractions that are favoured as is the case in barley (Kirkman et al. 1982). The influence of several other mineral nutrients on the composition of the storage protein fraction of pea seeds also were studied (Randall et al. 1979) and the effect of one of these, sulphur, has been characterized in detail.

The different classes of legume storage proteins have dissimilar sulphur-amino acid
contents although as a whole they are characteristically low in methionine and cysteine (see previous comments). The polypeptides of 7S globulin family are either low in or lacking these amino acids whereas those of the 11S (legumin-type) and the 2S (albumin-type) contain somewhat greater amounts (Schroeder 1984b; Huet et al. 1987). Although a limited sulphur supply marginally reduces the protein content of the seeds, the ratio of sulphur-containing to sulphur-free or sulphur-poor proteins is dramatically altered in both lupins (Gillespie et al. 1978) and peas (Randall et al. 1979). In peas, legumin is reduced to less than 5% of the amount present in the cotyledons of controls receiving adequate sulphur and there is a compensating increase in the level of vicilin (Chandler et al. 1983, 1984). These changes are correlated to changes in the levels of the various mRNAs for legumin and vicilin (Chandler et al. 1984). The effect of reduced sulphur supply on the levels of storage protein mRNAs can be reversed by addition of sulphur to either the whole plant, detached seed or isolated cotyledon (Chandler et al. 1984; T.J. Higgins, personal communication). This was investigated further by run-off transcription experiments with nuclei isolated from pea cotyledons of either sulphur-deficient or sulphur-restored plants (Beach et al. 1985). These in vitro transcription experiments showed that the changing levels of vicilin mRNA were regulated by changes to the level of transcription of vicilin genes whereas legumin mRNA accumulation was controlled by post-transcriptional events.

The supply of sulphur amino acids, specifically methionine, to developing legume cotyledons also was shown to alter the ratio of β-conglycinin to glycinin (legumin-type) in cultured soya bean cotyledons in favour of the 11S globulin (Thompson et al. 1981; Holowach et al. 1984a, b). No other amino acid when added to the culture medium induced similar changes. The reduction in the synthesis of β-conglycinin was restricted to the methionine-free β-subunit whereas those of the methionine-containing α and α' subunits were unaffected. These changes in the make-up of the globulin fraction were reflected in changes to the levels of the corresponding mRNAs (Creason et al. 1983). It can be concluded from these experiments that alterations to the supply of sulphur to the developing seeds affects the ratio between sulphur-containing and sulphur-lacking polypeptides. At least in one example (pea vicilin) this effect was shown to be regulated at the transcriptional level.
Phytohormones, particularly abscisic acid (ABA) have been shown to affect the level of storage proteins made by both cultured embryos and intact pods. ABA is known to inhibit precocious germination of embryos detached from the seed during the period of cell expansion and taken into organ culture. In experiments with seeds from soya beans and french beans, 10^{-5} M ABA in the nutrient solution prevented germination of the isolated embryos and maintained normal development including storage protein synthesis (Sussex and Dale 1979; Ackerson 1984 a, b; Eisenberg and Mascarenhas 1985). Production of storage proteins was markedly decreased at lower ABA concentrations. As was the case with the effects observed with sulphur-stress, these changes were reflected in the levels of specific storage protein mRNAs.

The levels of some storage proteins and their mRNAs are affected by prolonged culturing in the presence of ABA. For example, the level of mRNA for the β subunit of β-conglycinin was increased three-fold whereas those of the α and α' polypeptides were little changed (Eisenberg and Mascarenhas 1985). The ABA induced changes in β subunit mRNA could be inhibited by fluridone (Bray and Beachy 1985). Fluridone is an inhibitor of carotenoid biosynthesis and decreases the amount of ABA in cultured soya bean cotyledons. This inhibition was overcome by addition of more ABA to fluridone-treated cotyledons.

Although no effect of ABA could be observed with cultured pea cotyledons isolated from seeds at different stages of development (Davies and Bedford 1982), repeated injections of different phytohormones (including ABA) directly into the pedicles of pea pods over the course of seed development led to significant changes in the accumulation of storage proteins (Schroeder 1984 b). The three phytohormones used, naphthalene acetic acid (NAA, an auxin) benzyladenine (BA, a cytokinin) and ABA, all increased the total protein content per cotyledon. ABA induced an increase in vicilin accumulation whereas NAA and BA both increased legumin accumulation per cotyledon. When NAA and BA were injected together, an even more marked increase in legumin was observed but there was a decrease in vicilin content.
This section has reviewed experiments showing that the synthesis of some storage proteins can be affected by either changes to the level of mineral nutrients or by exogenous application of phytohormones and amino acids. In some cases, these effects are only observed in a particular class of storage protein or on a particular subunit within that class. Although the role of endogenous phytohormone in storage protein was not addressed by the experiments reviewed here, they do demonstrate that at least some storage proteins are able to respond to changes in hormone levels and suggest that these molecules may act as regulators in vivo.

7S globulin gene structure and chromosomal location.

7S globulins are encoded by small families of closely-related genes (Ellis et al. 1986; Schuler et al. 1982a; Talbot et al. 1984). In the case of pea and soya bean, the differences between the genes for the various subunits is sufficiently large that they can be distinguished and estimates of the number of genes present for each kind of polypeptide are available. So far this has not been possible for phaseolin as the α and β subunits are highly similar (see previous comments). Investigations on the number of genes encoding pea vicilin subunits have been complemented by linkage analysis studies that have allowed chromosomal locations for some of these genes to be determined (Matta and Gatehouse 1982; Ellis et al. 1986).

In pea, up to 18 genes are thought to encode the vicilin subunits of Mr 50,000 (Higgins 1984; Casey et al. 1986). Each of the classes of cDNA for these polypeptides (see previous comments) appears to be equally represented (Domoney and Casey 1985; Ellis et al. 1986). This finding is supported by evidence that the three classes of cDNA are transcribed at roughly equivalent levels (Beach et al. 1985).

Five genetic loci have been identified by Southern hybridization for the Mr 50,000 polypeptides (Ellis et al. 1986) with three loci found clustered on chromosome 7. Each locus contained a number of copies of closely related sequences with the sequence homologies within a locus being greater than those between loci. Previously Thomson and Schroeder (1978) had identified three vicilin loci designated Vca, Vcb and Vcc. The Vca locus was
identified because of variation in the convicilin polypeptide and will be dealt with later. The Vcb locus was identified because of variation in a Mr 30,000 polypeptide and the Vcc locus because of variation in the Mr 12,000 and 14,000 polypeptides. The observations of Thomson and Schroeder showed that these polypeptides behaved as the products of single loci with simple Mendelian inheritance patterns. However, in their paper, Ellis et al. (1986) argue that the Vcb and Vcc loci are not distinct but identical to one of the loci defined by them. This conclusion is questionable as it was assumed by the authors that the Mr 12,000, 14,000 and 30,000 subunits are derived by the complete processing of a single Mr 50,000 vicilin precursor. Experiments with expression of a pea vicilin gene in transgenic tobacco have shown that the mature seed can accumulate both processed and unprocessed forms of the introduced protein, indicating that processing can produce a number of polypeptides from a single gene product (this thesis, appendix 1; see previous comments). What is certain is that the genes encoding Mr 50,000 vicilin subunits are found at a number of distinct loci spread throughout the pea genome, a situation different to that found in soya bean and french bean where only one locus for the equivalent subunits exist (see below).

The number of genes encoding convicilin is uncertain and in part stems from its original definition as a protein of Mr 70,000 that is distinct from vicilin (Croy et al. 1980). As was mentioned previously, a single locus has been defined for this polypeptide (Vca in Thomson and Schroeder 1978; Cvc in Matta and Gatehouse 1982). This locus was found on chromosome 2. However, probing Southern blots of restricted genomic DNA from pea with cloned convicilin sequences detects two major hybridizing bands (Domoney and Casey 1985; Ellis et al. 1986; this thesis, chapter 3). Sequencing of genomic clones corresponding to these bands has demonstrated that they each contain a single copy of a gene similar to a number of cloned convicilin cDNAs (Bown et al. 1988; this thesis, chapter 3). One of these genes (but not the other) can be shown to be a likely candidate to encode Mr 70,000 convicilin (this thesis, chapter 3). However, the other gene is functional (Bown et al. 1988) and presumably gives rise to an as yet unidentified polypeptide similar in sequence but smaller than Mr 70,000 convicilin. It is obvious that the original definition of convicilin by Croy et al. (1980) is too
narrow in the light of recent data and should be broadened to include all the members of this family.

Uncertainties regarding the number of 7S storage protein genes also exist for soya bean. Davies et al. (1985) have identified 3 loci for β-conglycinin subunits based on differences in electrophoretic mobility. The loci for the α and β polypeptides (named Cgy₂ and Cgy₃, respectively) are genetically linked, although the linkage is not so tight as to preclude some recombination. The α' polypeptide locus (Cgy₁) does not seem to be linked to Cgy₂ and Cgy₃. Davies et al. (1985) concluded on the bases of a number of crosses between soya bean lines carrying different electrophoretic variants of the three subunits that there was only one functional coding sequence for each of the α and α' subunits, and two or more sequences for the β polypeptides. These findings are consistent with the estimate of 5 to 20 copies per genome of the 7S subunit genes reported by Goldberg et al. (1981) in liquid hybridization experiments using probes common to the α, α' and β-subunit mRNAs. The results for the α' subunit were in agreement with those obtained by studying a soya bean cultivar, Keburi, which lacks this polypeptide (Kitamura et al. 1984). The α' null allele in cv. Keburi is the result of a deletion of sequences from the 5' end and coding regions of the gene (Ladin et al. 1984). Furthermore, Southern blots of soya bean genomic DNA suggest that there are 8 to 13 copies of the β subunit gene per haploid genome (Tierney et al. 1987). However, Schuler et al. (1982 a, b) isolated several cDNA clones for the α and α' subunits which appears to disagree with the suggestion of Davies et al. (1985) that these subunits are encoded by single genes. This apparent contradiction between the genetic and the sequencing data is resolved by the possibility that the numerous cDNAs are allelic rather than the products of multi-gene families (R.N. Beachy, personal communication).

In the french bean, analysis of crosses have shown that there was only one locus for both the α and β polypeptides (Brown et al. 1981 a, b). This locus contained seven to nine copies of the genes for phaseolin (Talbot et al. 1984; Slightom et al. 1985).

Genes coding for representatives of all classes of 7S subunits mentioned in this introduction
have been isolated and their sequences determined (Doyle et al. 1986; Tierney et al. 1987; Bown et al. 1988; Weschke et al. 1988; this thesis, chapter 3 and appendix 1). The overall sequence organization of these genes is very similar as would be expected for genes of a common evolutionary origin. The coding regions are interrupted by five introns with the intron/exon boundaries obeying the GT/AG rule outlined by Breathnach et al. (1981).

Comparison of the gene sequences show that the introns are located in analogous places (Doyle et al. 1986; this thesis, appendix 1). Although there are similarities between the introns of β-phaseolin and α', β-conglycinin, none existed between those of vicilin and β-phaseolin. The coding regions of the genes are clearly similar but differ by both point mutations and deletions and duplications.

**Thesis outline**

In this introduction, I have reviewed aspects of the biosynthesis and regulation of the 7S storage proteins of legumes. It is apparent that 7S proteins react to a wide range of regulatory signals that limit their synthesis to particular times of seed development and cause them to respond to a number of environmental or hormonal changes. The major objective of the research reported in this thesis was to isolate two 7S storage protein genes from pea seeds and to determine both the primary DNA sequence of the genes and the characteristics of their expression in a new host species (tobacco). Of particular interest was the acquisition of sequence from the 5' flanking regions of these genes for comparison with the upstream regions of other related genes with the aim of defining stretches of conserved nucleotides that might be involved in regulating gene expression in an organ- and tissue-specific manner.

Initially, representative genes coding for members of both classes of pea vicilin, (Mr 50,000 vicilin and convicilin), were isolated from a genomic library and characterized by DNA sequencing. The sequences of these genes are reported in chapter 3 (convicilin) and appendix 1 (Mr 50,000 vicilin). When this research was initiated, the sequences of only two 7S genes had been determined (Schuler et al. 1982a; Slightom et al. 1983) neither of which was from pea.
The second step in this process was to characterize the expression of two isolated pea genes (one for Mr 50,000 vicilin and the other for convicilin) in transgenic tobacco. Results of these experiments are presented in chapter 3 and appendix 1. Previously, it had been shown that when a gene for β-phaseolin was transferred to tobacco, the features of its expression in the transgenic host were similar to those seen in french bean (Sengupta-Gopalan et al. 1985) indicating that the elements that regulated its expression were inherent in the DNA sequence. Expression of the pea genes in tobacco was used to demonstrate the presence of analogous regulatory elements in these genes.

Studies with a number of different genes have shown that transcription requires the presence of sequence elements upstream of the coding region (see Maniatis et al. 1987 for review). Some of these sequence elements are common to many genes whereas others are found only in the upstream regions of functionally-related genes (Davidson et al. 1983). To identify in the 7S gene family elements of potential regulatory significance, the sequences of the upstream regions from four 7S genes (the two pea genes and one each from soya bean and french bean) were compared. Results of these comparisons are given in Chapter 2. All of these genes were expressed in transgenic plants in a manner similar to that found in their original hosts and hence contained the regulatory element(s) of interest.

Comparisons of the upstream regions of the four 7S genes helped to identify short lengths of conserved sequence within 200 bp of the start of transcription that may act as regulators of gene activity. This area was thought to contain potential regulatory elements. As a final stage of the thesis, the region of conserved upstream sequence from the pea vicilin gene was isolated and used to direct the expression of a reporter gene (Chloramphenicol Acetyl Transferase) in transgenic plants. The results of these experiments are described in chapter 4.
Chapter 2

Mapping of the transcription start-points of two pea seed storage protein genes and identification of potential regulatory sequences in the 5' flanking sequence.
Introduction

Understanding the mechanisms by which transcription of eukaryotic protein-encoding genes are regulated is a central problem of molecular biology. Transcription requires the presence of sequence elements upstream of the coding region, some of which, such as the TATA and CAAT boxes, are common to many genes (Dynan and Tjian 1985; Maniatis et al. 1987). Apart from these elements, short (<100bp) sequences have been identified that are found only in the upstream regions of functionally related genes (Davidson et al. 1983). It has been suggested that these sequences have a role in the regulation of transcription of such genes. In some cases, direct evidence has been obtained, by in vivo expression studies, that these sequences are required for correctly regulating transcription. Upstream elements have been found in several plant promoters including those responsive to light (Fluhr et al. 1986; Simpson et al. 1986), heat shock (Schoffl et al. 1986), anaerobiosis (Ellis et al. 1987), embryo development (Chen et al. 1986) and to nodulation (Stougaard et al. 1987).

The upstream regions of four 7S storage protein genes are compared in this chapter to help identify conserved sequences that may act as regulatory elements. The four genes chosen encode pea vicilin, pea convicilin, soybean α', ß conglycinin and ß-phaseolin from French bean. The sequences of the two pea genes are reported elsewhere in this thesis (Chap. 3 and appendix 1) and the other two sequences are both taken from Doyle et al. (1986). These genes were selected not only because a considerable length (at least 1kb) of upstream region had been determined and was available for comparison but also because all the genes have been expressed in transgenic plants (see appendix 1 and Chap. 3 of this thesis; Beachy et al. 1985; Sengupta-Gopalan et al. 1985) and therefore have the potential to act as functional genes in their original hosts.

In this chapter, the location of the start sites of transcription of the two pea genes was determined and the 5' flanking sequences of the four storage protein genes were compared. The upstream sequences were found to be largely dissimilar except for short sequences just upstream of the start points of transcription that were common. These conserved sequences
were identified and experiments intended to determine their role in regulation of expression are discussed.

Methods

**DNA sequencing and sequence comparisons.**

The 5' flanking sequences of vicilin, convicilin, β-phaseolin and α', β-conglycinin were compared using version 9.2 of the Diagon programme (Staden 1982). Isolation of the two 7S storage protein genes from pea and determination of their DNA sequence was done as described elsewhere in this thesis (Chap.3 and appendix 1). The sequences of the other two genes are taken from Doyle *et al.* (1986). The Diagon programme produces a two dimensional plot that is a representation of all the matches in a pair of sequences. The X-axis represents one sequence and the Y-axis the other. Both sequences are positioned such that their 5' ends are at the bottom left hand corner. The proportional matching algorithm used in this chapter counts the number of identical pairs of bases at the same position in a given length of sequence called the span. When this number equals or exceeds a given value called the score, a dot is placed on the plot at a position corresponding to the centre of the span. Both the length of the span and the score are defined by the user. The span is moved along the sequences one nucleotide at a time until all positions on the plot have been considered.

**S1 mapping of vicilin and convicilin transcripts from pea seeds.**

The start sites of transcription of the vicilin and convicilin genes were mapped using the S1 nuclease protection assay. This was performed as described by Weaver and Weissman (1979) using single-stranded DNA probes (Hudson *et al.* 1985). The two probes to be used were DNA fragments that had been ligated into M13-based plasmids resulting in clones called p220 and C3. The inserts in p220 and C3 were complementary to the mRNAs encoded by pEN2 (vicilin) and pEN6 (convicilin), respectively. The p220 insert covered the first 143 bases and the C3 insert the first 169 bases of the coding regions of the genes from which they were derived. Both inserts included at least a further 800 bases upstream of the coding region. Single-stranded phage DNA, prepared from JM101 cells infected with either p220 or C3, was
purified on cesium chloride gradients as described by Grundstrom et al. (1985). Radioactive, single-stranded probe was made by annealing 2.5ng of M13 universal sequencing primer (17-mer, BRESA, South Australia) to approximately 0.8μg of purified phage DNA in TM buffer (0.1 MTris/Cl pH8.5, 50mM MgCl2) for 20 minutes at 55°C. The template was extended with the Klenow fragment of DNA polymerase I (BRESA, South Australia) using 0.05mM each of dATP, dTTP, dGTP and 20μCi of dCT32P (BRESA, South Australia) for 10 minutes at room temperature. The reaction was then chased for a further 10 minutes with cold deoxynucleotides. The enzyme was heat inactivated at 70°C for 15 minutes and NaCl was added to the solution to a final concentration of 100mM. Five units of EcoR1 restriction enzyme (Boehringer, Mannheim) were added and the reaction incubated at 37°C for sixty minutes. Formamide dye marker (Bankier and Barrell 1983) was added and the reaction boiled for 3 minutes prior to electrophoresis on a 5% denaturing acrylamide gel. Bands were recovered from the gel by the method of Maxam and Gilbert (1980).

Total seed RNA was extracted from cotyledons harvested 17 days after flowering (DAF) by the method of Chandler et al. (1983). S1 mapping was done by mixing 50-100,000 cpm of freshly-made probe with 1μg of total pea seed RNA and 50μg of transfer RNA. This was precipitated with ethanol and redissolved in 15μl of hybridization buffer (80% formamide, 0.5M NaCl, 40mM PIPES pH6.4, 1mM EDTA) heated to 60°C for 10 minutes and allowed to hybridize for 14 to 16 hours at 37°C. The samples were then diluted with 400μl of cold, S1 buffer (30mM sodium acetate pH4.5, 0.25M NaCl, 1mM ZnSO4, 5% glycerol) containing 400 units of S1 nuclease (Boehringer, Mannheim) and incubated at 37°C for 45 minutes. The reaction was terminated by adding 100μl of stop solution (4M ammonium acetate, 0.1mM EDTA, 10μg transfer RNA) and precipitated with an equal volume of isopropanol. After centrifugation, the pellet was dissolved in formamide dye, boiled for 3 minutes and electrophoresed on a 5% denaturing acrylamide gel. The gel was subsequently dried and exposed to film at -70°C using a Hi-plus (Dupont) intensifier for 16 hours.
Results

The sequences of pEN2, which contains a gene for a pea vicilin, and pEN6, a clone that includes a portion of the pea convicilin gene, are presented elsewhere in this thesis (appendix 1 and chapter 3, respectively) and only the 5' flanking regions of these genes will be considered here. About 1,500 bp 5' of the coding region of pEN2 and nearly 1,100 bp of the upstream sequence of pEN6 were determined. The start points of transcription of these two genes were defined by S1 nuclease protection assay as described in the Methods section. As probes, uniformly labelled, single-stranded DNA molecules were used. The probes were complementary to around 150 to 170 bases of the amino-terminal encoding region of the mRNA derived from either the vicilin or convicilin gene and had at least a further 800 bases upstream of this region. The results of such an experiment are shown in Figure 1a. In both cases, more than one protected fragment was observed indicating that transcription was initiated at a number of sites in these genes. For the vicilin gene (pEN2) two protected fragments, 180 and 183 bases in size, were observed whereas for the convicilin gene (pEN6), four bands of 197, 202, 206 and 210 bases were seen (Fig.1a). In figure 1b, the alignment of these bands with the nucleotide sequences of the genes is shown. The base designated +1 was the most intense band in each case.

The 5' flanking regions of four, 7S storage protein genes were compared using the proportional matching algorithm of the Diagon programme (Staden 1982). Apart from the two pea genes presented here, the sequences of a β-phaseolin gene from french bean and the α', β-conglycinin gene of soy bean were also studied (Doyle et al. 1986). For the purposes of these comparisons, the 950 bp immediately 5' of the coding region of each gene were used.

Figure 2 shows the effect on a comparison of convicilin (X-axis) and vicilin (Y-axis) 5' regions of different settings for the span or score (see methods). Since with nucleic acid sequences around 25% of the span will contain pairs of identical nucleotides, care has to be taken in selecting both the span length and the score to enable the significant similarities to be distinguished from the chance matches. As the span length was increased from 9, through 15,
Figure 1a: Determination of the start site of transcription for two 7S genes, vicilin (A) and convicilin (B), from pea by S1 mapping. These experiments were done as described in the methods section using 1μg of total RNA isolated from pea seeds harvested 17 days after flowering. The numbers on the vertical axis give the size in nucleotides of standard DNA fragments (PBR322 digested with Msp1).

Figure 1b: The location of the start sites of transcription on the nucleotide sequences of vicilin (upper line) and convicilin (lower line). In both cases, the base numbered +1 was the most intense band seen in figure 1a. Alternative start sites are indicated by an asterisk (*). The underlined nucleotides at the 5' ends of the sequences are the presumed TATA boxes. The initiator methionine codon (ATG), which marks the beginning of the reading frame, is at the 3' end of the sequences.
Vicilin (pEN2).

```
CTATAATACACTCATCTCAGCATCCTTTATTTCCACCGAAGAGGATGGATAGTTTCAATCAACATG
```

Convicilin (p5.1).

```
CTATAATATTCTCATCTCTTTTATTTTCCATCTCAAACCTTTTAACTGAAATACAAACTCATG
```
to 21 bases (Fig.2, panels a to c), the number of dots resulting from chance matches decreased without affecting the extent of significant homologies (seen as a broken diagonal line of dots in the top right hand corner of the plots in Fig.2). The percentage of matching nucleotides (score) in all these cases was between 71 and 78%. Furthermore, when the score was decreased from 71% to 61% and 51% (Fig.2, panels c d and e) with the span set at 21, the background increased markedly with little effect on the level of homology. Since the line of homology was more evident when the span was set at 21 bases and the score at 71%, these settings were used for the remainder of the comparisons.

Figure 3 shows the Diagon plots resulting from the pair-wise comparisons of the 5' flanking regions of the four 7S storage protein genes mentioned earlier. The homology between any two sequences was limited in all cases to regions near the 3' ends of these sequences (ie. proximal to the coding region). The scatter of dots seen in the lower left quarter of some of these plots (particularly Fig.3, panels d, e and f) resulted from the 5' ends of these sequences having a higher percentage of "A" or "T" nucleotides than the remainder which increased the likelihood of chance matching. For example, overall the α', β-conglycinin sequence used here was 32% "T" and 35% "A" nucleotides, however within the first 300 bases at the 5' end the sequence it was 44% "T" and 39% "A" nucleotides. It seems probable that these dispersed points did not result from significant sequence conservation in areas distant from the coding region.

The similarity between any two genes was most extensive for comparisons between convicilin and α', β-conglycinin (Fig.3b) and between α', β-conglycinin and β-phaseolin (Fig.3f) where the line of homology commences about 300-350 bases from the coding region in all sequences. The pair of genes that were least similar were β-phaseolin and vicilin (Fig.3d) where only a single short region of homology starting about 100 bp upstream of the coding region was observed.

The exact locations of all homologous regions on the nucleotide sequences were determined using the cross-hair option of the Diagon programme. Although there were many sequences
Figure 2: Diagon plots of a comparison between the upstream sequences of the genes for a pea vicilin (y-axis) and pea convicilin (x-axis) showing the effect of changes to the span length or the score. The score for panels a, b and c was between 71 and 78% with span lengths of 9, 15 and 21 nucleotides, respectively. The span length for panels d and e was 21 with the scores set at 61 and 51%, respectively.
Figure 3: Diagon plots for the comparisons of upstream regions of the following pairs of genes.

Panel a: convicilin and vicilin; panel b: convicilin and α', β-conglycinin; panel c: convicilin and β-phaseolin; panel d: vicilin and β-phaseolin; panel e: vicilin and α', β-conglycinin; panel f: β-phaseolin and α', β-conglycinin. In each case the first mentioned of the pair of sequences is the x-axis. For each plot the span was set at 21 and the score at 71%.
that were common to two genes, only two regions were found to be shared by three or more sequences. The length and position of these two regions on each upstream sequence is shown in Fig.4. The longest conserved region (named region I) was about 70 bp long and started at approximately -130 in the four genes. Region I was somewhat shorter in β-phaseolin (about 40 bp). A second region (named region II) of about 20 bp was common to three of the genes (vicilin, convicilin and α', β-conglycinin) although its position in these three genes was not identical. In vicilin, this region was adjacent to region I whereas in convicilin and α', β-conglycinin, it was approximately 150 and 190 bp further upstream, respectively.

Figure 5 shows the vicilin 5' flanking region compared at the nucleotide level to the other three flanking sequences and aligned for maximal identity. Numbering in this figure is relative to the start of transcription of vicilin (Fig.5, line a) and ignores padding characters. A stretch of 21 identical nucleotides common to vicilin and convicilin was observed starting at -184 (numbering from the 5' end). Nineteen of these 21 bases were also found in α', β-conglycinin although the similarity to vicilin continued further upstream than it did in convicilin. This sequence, which was not observed in β-phaseolin, corresponded to region II in Fig.4.

The region of shared sequence which corresponded to region I started at position -131 with the hexanucleotide "GCCAC(C/T)". This was followed by about forty nucleotides which were approximately 76% conserved between the four genes. Due to an extra 6 bases in β-phaseolin, a gap was introduced in the other three flanking sequences at position -83. Another sequence of 11 common nucleotides was found at position -80. A part of this region was similar to the consensus sequence "GG(C/T)CAATCT" or CAAT-box found in many eukaryotic genes transcribed by RNA-polymerase II. A sequence of eight conserved nucleotides was seen just downstream of the presumptive CAAT-box starting at position -66. This marked the 3' end of region I.

Two short sequences, not seen in Fig.4 were observed downstream of region I. A conserved sequence with homology to the TATA-box was found at -33 and the nucleotides "TTCATC" were conserved at a position at or near the start point of transcription in all four flanking
Figure 4: Diagram showing the location of conserved sequences in the 5' flanking regions of four 7S storage globulins. Line A is vicilin; line B is convicilin; line C is α', β-conglycinin; line D is β-phaseolin. The conserved regions were identified as outlined in the text. Sequences outside the boxed regions were less than 70% identical. Numbers under the lines were determined from the transcription start site of each gene.
regions. Downstream of this point are the 5' untranslated leader sequences of the various messenger RNAs. Neither the length nor the nucleotide sequence of this region were conserved.

Discussion

This chapter presents a detailed comparison of the 5' flanking sequences of four 7S seed storage protein genes. This was done on the assumption that the elements important in gene expression are likely to be conserved among a group of genes with the same pattern of development (Davidson et al. 1983).

The four genes studied here are presumed to be derived from common ancestral genes because of the similarity of their coding regions (Argos et al. 1985; Borroto and Dure 1987). This similarity may reflect constraints on the mutability of these proteins due to functional requirements for post-translational processing (including the formation of oligomers); correct targeting to the protein body via the Golgi (Chrispeels 1985) or the efficient packing of oligomers within the protein body (Pernollet and Mosse 1983). Despite the conservation within the coding regions, there were only short sequence elements that were common to the four genes in their 5' flanking regions (Fig. 5). Apart from these sequences, the upstream regions were divergent.

One of the short sequences that was shared by these genes was identical to the TATA box consensus "TATAAA" (Proudfoot 1979). TATA boxes are mechanical elements of the promoter needed for correct initiation of mRNA transcription, typically at a point 25-35 bp further downstream (for review, see Breathnach and Chambon 1981). In agreement with this, S1 mapping showed that the major transcripts derived from the pEN2 (a pea vicilin gene) and pEN6 (a pea convicilin gene) began 27 and 33 bp, respectively, downstream of the TATA box (Fig. 1). Other minor transcripts were also observed. Variability in transcription initiation is not unusual (Hoffman 1984) and generally involves only two to three nucleotides near the cap site. Six nucleotides were conserved in all four genes at or near the cap site although the length
and sequence of the untranslated leaders were dissimilar (Fig. 5).

Upstream of the TATA box, a typical promoter has one or more short sequence elements that are required for efficient and regulated transcription (Davidson et al. 1983; Dynan and Tjian 1985; Maniatis et al. 1987). One of these upstream elements, the CAAT box is found in many different promoters. Single base pair changes within the CAAT box of a β-globin gene markedly affected the ability of this promoter to direct transcription of the adjacent coding region (Myers et al. 1986). A sequence beginning at -79 in pEN2 and shared with the other 7S storage protein promoters studied shows similarity to the consensus sequence for this element (Benoist et al. 1980; see Fig. 5).

Although the TATA and CAAT boxes are common to many different promoters, they are not sufficient to initiate transcription. Further upstream elements are required to impart to the promoter characteristics that are typical of its normal expression. Short upstream sequences, conserved among a number of functionally related genes have been shown to be involved in the regulation of these genes (Davidson et al. 1983). Among the plant genes so far studied, comparison of the promoter sequences of co-ordinately-induced genes such as the heat shock (Schoffl et al. 1986) or light-regulated genes (Fluhr et al. 1986; Simpson et al. 1986) has shown sequences similarities. Furthermore, in both cases, direct evidence has been obtained by DNA transformation studies that the conserved regions are involved in regulation (Baumann et al. 1987; Fluhr et al. 1986).

The location of the upstream regions conserved among four 7S storage protein genes is shown in figure 4 with the direct sequence comparisons given in figure 5. Two regions, apart from the CAAT and TATA boxes and the homology near the cap site, were observed that were shared among at least three of the four genes (Fig. 5). Region I, lying between -131 and -59 in pEN2 (vicilin), began with the sequence GCCACGT and contained the CAAT box. A part of this region has been described as the vicilin box (from approximately position -132 to -90; Gatehouse et al. 1986). The other conserved region (region II) was between -183 and -164 in pEN2 and was not found in the phaseolin sequence.
Experiments in transgenic plants with deletions of the upstream regions of the α' subunit of β-conglycinin have defined two areas that control the embryo-specific and high level expression of this gene (Chen et al. 1986). Deletion of sequences from the 5' end to a point 257 bp from the cap site reduced neither the level of transcription nor its specificity. Deletion of a further 40 bp to -208 reduced the level of transcription by a third. Removal of a further 50 bp to -159 decreased the level of transcription to 5% of that found with the full (8.5kb) 5' flanking sequence, however the specificity of this expression was unaffected. This deletion leaves the vicilin box intact. No expression could be measured from conglycinin genes that had been deleted to a point 69 bp upstream of the cap site (which would leave both the CAAT and TATA boxes) or beyond. Furthermore, when a 170 bp sequence (-257 to -78) of the conglycinin promoter was placed in different positions and orientations into a construct comprised of a strong constitutive plant promoter, the 35S promoter from cauliflower mosaic virus (CaMV), and a reporter gene (cat), there was an enhancement of cat gene expression in the seeds of transgenic tobacco plants above that which was observed in plants transformed with CaMV 35S promoter and cat alone (Chen et al. 1988). This effect was largely independent of the orientation or location of the conglycinin promoter fragment within the construct and indicates that there was a sequence (or sequences) within the fragment that acted as a seed-specific enhancer.

Taken together, the identification of conserved sequences in the upstream regions of four 7S storage protein genes and experiments on the promoter of one of these genes (Chen et al. 1986; Chen et al. 1988) indicate that sequences sufficient for seed-specific expression of these genes may lie close (within ~250 bp) to the cap site. Among the many random fragments produced in the sequencing of pEN2, one (p373) that covered the region from -227 to -49 included all the conserved regions shown in figure 5 (except for the TATA box and the cap site) This fragment was tested for its ability to direct seed expression of a reporter gene and the results of these experiments are presented in chapter 4.
Chapter 3

Pea convicilin: Structure and primary sequence of the protein and its expression in the seeds of transgenic tobacco

Edward J. Newbigin1,5, Ben O.deLumen2, Peter M. Chandler1, Alan Gould3, Robert J. Blagrove4, John March6, Alex Kortt4 and Thomas J.V. Higgins1

Present address:
1; CSIRO Division of Plant Industry, GPO Box 1600, Canberra, Australia 2601.
2; Department of Nutritional Sciences, University of California, Berkley, California U.S.A. 94720.
3; United Agriseeds, Inc., PO Box 4011, Champaign Illinois, U.S.A.
4; CSIRO Division of Biotechnology, 343 Royal Pde, Parkville 3052 Australia.
5; Department of Botany, Australian National University, GPO Box 4, Canberra, 2601 Australia.
6; John Innes Institute, Coloney Lane Norwich NR4 7UH United Kingdom.

Key words: Convicilin, transgenic tobacco, Agrobacterium tumefaciens
Abstract

Convicilin, a trimeric globulin of pea seeds, is closely related to vicilin and composed of polypeptides of Mr~70,000 kilodaltons (kd). A partial cDNA clone encoding convicilin was isolated, sequenced and used to select a convicilin gene from a pea genomic library. A part of the genomic clone was sequenced to obtain the coding sequences missing in the cDNA clone and a further 1 kilobases (kb) 5' to the start of transcription were also obtained. The entire sequence of convicilin was deduced from the combined genomic and cDNA sequences. The complete gene encoding convicilin was transferred to tobacco and the characteristics of its expression in the seeds of transgenic plants were studied. A protein, not observed in the seeds of untransformed tobacco, was found only in the seeds of the transgenic plants. The protein was identical in size to pea convicilin and recognized by anti-vicilin antiserum. The expression characteristics and the transcription start sites of the gene in tobacco were similar to those observed in pea. Convicilin, which does not undergo post-translational cleavage in peas, was partially processed to polypeptides of Mr~50,000 in tobacco seeds. There was a twofold variation in the level of convicilin accumulated by the mature seeds of a number of transgenic plants and this was well correlated with the number of gene copies incorporated in the different transformants. In the seeds of tobacco plants that contained a single copy of the transferred gene, it was estimated that convicilin comprised about 2% of the seed protein.
Vicilin constitutes a major fraction of the protein in a pea seed. Native vicilin is a trimeric 7S protein that contains a number of polypeptides, some of which are glycosylated (for review, see Casey et al. 1986). Vicilin can be divided into two classes of polypeptides on the basis of the size of the primary translation product from which the polypeptides are derived. Those of sizes of Mr 50,000 or less result from cleavage after translation of a family of Mr 50,000 vicilins at one, both or neither of two internal processing sites (Gatehouse et al. 1983; Spencer et al. 1983). Some members of this class are also glycosylated (Badenoch-Jones et al. 1981; Chrispeels et al. 1982b). The second class, containing only polypeptides of Mr~70,000, has been called convicilin (Croy et al. 1980). Convicilin is not glycosylated (Croy et al. 1980) and does not appear to undergo any further processing such as endoproteolytic cleavage (Chrispeels et al. 1982a). The two classes of vicilin are antigenically related (Croy et al. 1980) but encoded by distinct gene families that are not genetically linked (Ellis et al. 1986) and are active at different times of seed development (Chandler et al. 1984). The Mr 50,000 vicilins are encoded by approximately 18 genes (P.M. Chandler, unpublished observation) that are transcribed early in the deposition of storage reserves. Convicilin, on the other hand, is encoded by a single gene (Domoney and Casey 1985; Ellis et al. 1986) although subsequent data indicates the presence of at least one other related sequence (Bown et al. 1988; this paper). Both of these genes are active during the later stages of seed development (Chandler et al. 1984; Bown et al. 1988).

When this work was initiated, the precise relationship between convicilin and vicilin was not known and there was some uncertainty about the oligomeric nature of convicilin (Croy et al. 1980; Casey and Sanger 1980). In this paper, we report the oligomeric structure of convicilin, the complete primary sequence, the isolation of the convicilin gene and features of its expression when transferred to tobacco.
Methods

Isolation of a convicilin gene and sequencing of cDNA and genomic fragments

A genomic library of pea, constructed in λ1059 (Higgins et al., 1988), was screened with the cDNA pPS15-28 that had earlier been characterized as encoding a 68.2 kd (previously called Mr 75,000) convicilin (Chandler 1982; Chandler et al. 1984). A single positive clone was purified and characterized by restriction mapping. A 7.5kb EcoR1 fragment that hybridized on Southern blots (Southern 1975) to pPS15-28 was subcloned into pUC8 (p5.1) for further study. An EcoR1 to Pst1 fragment (2.1kb) that hybridized to a part of pPS15-28 encoding the N-terminal end of the coding region was subsequently cloned into pUC8 (pEN6). The inserts of pEN6 and pPS15-28 were further characterized by sequencing using the dideoxy chain termination method (Bankier and Barrell, 1983) and the chemical sequencing method of Maxam and Gilbert (1980), respectively.

Purification, oligomeric structure and N-terminal sequence of convicilin

Convicilin was extracted from dried, mature seeds of Pisum sativum c.v. Birte by the method of Casey and Sanger (1980). The protein, was purified by a combination of (NH₄)₂SO₄ fractionation, zonal isoelectric precipitation and DEAE-cellulose chromatography (Casey 1979; Casey and Sanger 1980). Immediately prior to molecular weight measurements in the ultracentrifuge, the sample was fractionated by chromatography over Sepharose CL-6B.

Methods for ultracentrifuge analysis, measurement of circular dichromic spectra and chemical crosslinking with dithiobis (succinimidyl propionate) (DTSP) have been reported (Blagrove et al. 1984). N-terminal amino acid sequencing was carried out on an Applied Biosystems 470 A gas-phase sequencer. The isolated PTH-amino acids were identified by HPLC on a Zorbax ODS column.
Transfer of convicilin gene into tobacco

The plasmids BDL5 and pAG1 were made by ligating the 7.5kb EcoR1 fragment of p5.1 into the unique EcoR1 sites in vectors Bin19 (Bevan, 1984) and pGA492 (An, 1986), respectively. These plasmids were transferred from *E. coli* (HB101 or JM83) to *Agrobacterium tumefaciens* using the tri-parental mating method of Ditta *et al.* (1980). Transconjugants were selected on LB agar containing rifampicin (50μg/ml), streptomycin (200μg/ml) and tetracycline (5μg/ml) and the presence of the gene was confirmed by Southern blotting. Transformation of tobacco (Wisconsin 38) leaf pieces and regeneration of transgenic plants was done as described in Higgins *et al.* (1988). Flowers were tagged at anthesis.

Protein detection by Western blotting.

Total protein was isolated from fresh tissue as described in Higgins *et al.* (1988) and protein concentration was determined by the method of Bradford (1976). Proteins were electrophoretically separated either on 12.5-25% SDS-polyacrylamide gels as in Spencer *et al.* (1980) or on two-dimensional gels (adapted from the method of Hughes and Galau, 1985). The proteins were then electroblotted onto nitrocellulose membrane as described (Spencer *et al.*, 1983) and convicilin was detected using primary antibodies generated in rabbits against total pea vicilins (including convicilin) followed by a second antibody (goat, anti-rabbit IgG) linked to alkaline phosphatase (Bio-Rad, Richmond, CA) for colorimetric detection.

DNA isolation and Southern blotting of genomic DNA

DNA was isolated from young tobacco leaves or pea shoots by the method of Sutton (1974). After digestion with EcoR1 (Boehringer, Mannheim), the DNA was fractionated on 1% agarose gels and transferred to nylon membrane (Zeta Probe, Bio-Rad, Richmond, CA) prior to UV treatment as described in Khandjian (1987) except that the gel was washed in 0.25M HCl for 5 minutes before denaturation with NaOH. Hybridization was carried out using the 2.1kb insert of pEN6. This was labelled by nick translation using a kit (BRESA, Adelaide, South Australia).
Hybridizations were done as described by Khandjian (1987). The membranes were then given four fifteen-minute washes at 65°C in 2xSSC containing 0.1%SDS and 0.1% sodium pyrophosphate and then four fifteen-minute washes at 50°C in 0.1xSSC containing 0.1%SDS and 0.1% sodium pyrophosphate. Blots were then fluorographed at -70°C using a Hi-plus (Dupont) intensifier for four days.

RNA isolation, S1 mapping and densitometry

Total RNA was isolated from about 500mg of frozen tobacco seeds as described by Higgins et al. (1988). Total RNA was isolated from pea cotyledons as described in Chandler et al. (1983). S1 mapping was done as described (Weaver and Weissman, 1979) using a single-stranded probe (Hudson et al., 1985). The probe extended at least 800 bases upstream of position +197 (see Fig.2). The protected fragments were separated on urea/acrylamide gels using 32P-labelled Msp1 fragments of pBR322 as size markers.

Densitometry of fluorographs and negatives was done as previously described (Higgins et al., 1988).

Results

Isolation and structure of a pea convicilin gene

A cDNA (pPS15-28), previously characterized by hybrid-arrest translation as encoding a 68.2 kd convicilin polypeptide (formerly referred to as a Mr 75,000 polypeptide; Chandler 1982; Chandler et al. 1984), was used to screen a pea genomic library constructed in λ1059. A positive clone was mapped with restriction endonucleases and a 7.5kb EcoR1 fragment that hybridized to pPS15-28 was subcloned into pUC8 resulting in clone p5.1. The orientation of the gene in p5.1 was mapped by Southern blotting. This identified an EcoR1 to Pst1 fragment of 2.1kb as encoding the N-terminus of the protein and about 1kb of sequence 5' to the coding region. A restriction map of the genomic clone p5.1 showing the position of the gene is given
Figure 1. The restriction map of a 7.5kb EcoR1 fragment containing the gene for convicilin from pea. The upper line shows the full length insert in p5.1. The lower line shows the 2.1kb EcoR1 to Pst1 fragment (pEN6), which was sequenced (see Fig.2) with the coding region shown by the heavy line. The open region shows the position of the first intron. An arrow indicates the direction of transcription. The open arrow head shows the beginning of the convicilin cDNA clone, pPS15.28. The closed arrow head marks the position at which convicilin and vicilin begin to show sequence similarity. The letters E, P, N and X refer to the restriction enzymes, EcoR1, Pst1, Nco1 and Xba1, respectively.
in Fig.1. The 5' EcoR1 to Pst1 fragment (2.1kb) was subcloned into pUC8 and named pEN6. The sequences of pPS15-28 and pEN6 were both determined and are shown in Fig.2. The sequence of the encoded protein was deduced by the presence in the genomic clone of a long open reading frame with a methionine at its 5' end. The first 29 amino acids were followed by a sequence which was almost identical to the chemically determined N-terminal sequence NYDEGSEP(R)VPAQ(R)E of purified convicilin from pea seeds. The first 29 amino acids were characteristic of eucaryotic signal sequences (von Heinje 1983).

The sequence of the cDNA overlapped the genomic sequence starting at a position 313 bp downstream of the initiator methionine. The presence of an intron of 151 bp found 805 bp from the beginning of the reading frame in the genomic clone was inferred by comparison with the nucleotide sequence of the cDNA clone. There was an additional 84 bp past the end of the intron which were shared between the genomic and cDNA sequences before the Pst1 site which marks the end of the determined genomic sequence. Over the entire 576 bp common to the two clones, there were four nucleotide differences which resulted in two amino acid changes. Both the amino acid changes were from arginine in the genomic sequence to lysine in the cDNA. The genomic sequence was compared to another partial cDNA sequence for convicilin, pCD59 (Casey et al. 1984), and five differences were found in 321 bases compared (data not shown).

The genomic clone had 1046 bp of sequence 5' to the coding region. A "TATA" box (Proudfoot 1979) was found 66 bp upstream of the start of the coding region. The start point of transcription was determined by S1 nuclease protection. Four protected bands of sizes 197, 202, 206, and 210 bases were observed with the band of 197 bases being the most intense (see Fig.8, lane 8). The base at this position was numbered +1. A sequence was found at -83 which shows similarity to the "CAAT" box (Benoist et al. 1980) found in many animal and plant genes transcribed by RNA polymerase II. Upstream of the start of transcription, starting at -137 with the nucleotides GCCACC and continuing for about 40 bp 3' of this position, was a sequence named the "vicilin" box that has been observed in all the 7S storage protein genes
Figure 2. The nucleotide sequences of parts of the convicilin genomic and cDNA clones and the deduced amino acid sequence of convicilin. The genomic DNA sequence was determined by the method of Bankier and Barrell (1983). The cDNA sequence was determined by the method of Maxam and Gilbert (1980). Line a=genomic sequence of pEN6; b=convicilin cDNA sequence (pPS15-28); c=protein sequence deduced from pEN6; d=protein sequence deduced from pPS15-28. * = identity of bases or amino acids. The "vicilin" (v), "CAAT" (c), and "TATA" (t) boxes (-137, -83 and -38, respectively, numbering from the 5' end) are overlined. An arrow head indicates the base designated +1. The mature N-terminus of the protein is indicated with an arrow. The two amino acid repeats are underlined with vertical bars indicating the ends. The reiterated polyadenylation sequence, AATAAA, found in the 3' untranslated region of the cDNA are boxed.
from legumes so far examined (Gatehouse et al. 1986; Bown et al. 1988; Higgins et al. 1988). The nucleotides AATAAA were repeated six times between the translation termination site and the polyadenylation site at the end of the cDNA clone.

Convicilin sequence and structural comparisons with vicilin

The amino acid sequence deduced from the complete cDNA and genomic nucleotide sequences resulted in a single reading frame of 607 amino acids. Removal of the first 29 amino acids that presumably act as a signal peptide would give a polypeptide of 578 residues with a molecular weight of 68.2 kd. Starting at position 629 in figure 2, the C-terminal sequence of convicilin showed similarity to the Mr 50,000 vicilins, as has been previously noted (Casey et al. 1984; Casey et al. 1986; Bown et al. 1988). The N-terminal 166 amino acids are unique to convicilin and account for the size difference between the Mr 50,000 vicilins and convicilin. This region was highly hydrophilic with 60% of the residues being arginine, lysine or glutamate. It contained two direct duplications, one of 32 and the other of 15 amino acids. The larger duplication had three differences and the smaller had one difference at the amino acid level between the two occurrences. The second occurrence of the first duplication overlapped the first occurrence of the second duplication by five amino acids (see Fig.2).

The high ratio of charged to apolar residues in the N-terminal region of convicilin generally would lead to a rod-shaped domain consistent with the presence of an extended conformation (possibly α-helical). Structural predictions (Garnier et al. 1978) based on the deduced amino acid sequence of this region anticipates helix levels of 40 to 60% depending on the decision constants chosen for the analysis. If these predictions were accurate, convicilin should show a measurable increase in the helix content (from 5 to 10% to 15 to 20% α-helix) compared with the Mr 50,000 vicilin molecule (assuming that the conformation of the C-terminal vicilin-like region is not perturbed by the additional sequence).

Circular dichroism measurements at pH 7.0 (results not shown) gave near identical spectra
Figure 3. Electrophoresis on SDS polyacrylamide gels of convicilin (0.2 mg of protein dissolve in 0.1 ml of 0.05 M phosphate buffer pH 6.8) crosslinked with DTSP. (1) untreated control; (2) 0.01 mg DTSP; (3) 0.03 mg DTSP; (4) 0.1 mg DTSP; (5) 0.1 mg DTSP sample treated with 2% 2-mercaptoethanol before electrophoresis. Apparent molecular weights (Mr x 10^{-3}) are marked on the vertical axis.
for vicilin and convicilin indicative of proteins with a low helix content (less than 5% α-helix). In case the negative charge on the dominant glutamate residues was inhibiting helix formation (poly-L-glutamic acid requires partial protonation at pH 4 to induce helix formation), the circular dichroism measurements were repeated at pH 4.0 with no appreciable change in the spectra, especially in the helix-sensitive region of 208 to 222 nm. While the N-terminal extension of convicilin may well form an extended surface domain, this region was not α-helical in the manner predicted from amino acid sequence.

There is uncertainty in the literature as to the oligomeric form of native convicilin. Reports from different groups have shown that convicilin can be arranged either as a trimer (Casey and Sanger 1980) or as a tetramer (Croy et al. 1980) of Mr 70,000 subunits. Typically, 7S proteins are trimers (Casey et al. 1986). To clarify the quaternary structure of convicilin as extracted from mature seeds with aqueous salt solutions at neutral pH, the molecular weight of the native protein was determined by a meniscus-depletion sedimentation equilibrium experiment. The major component of convicilin in 0.5 M NaCl at pH 6.8 proved to be a homogeneous protein of Mr 190,000 ± 5,000, a value close to that determined by Casey and Sanger (1980) and indicating a possible trimeric structure. In order to confirm the trimeric nature of convicilin, chemical crosslinking experiments with DTSP were performed and the results analysed by SDS-PAGE (Fig.3). Addition of DTSP led to the appearance of two further broad bands of higher apparent molecular weight than convicilin itself. With increasing amounts of crosslinker, there was complete conversion to the largest species with little evidence of higher aggregates. The fully polymerized form could be reduced with 2-mercaptopoethanol to regain subunits with modified lysines. The results in figure 3 are similar to those reported (Blagrove et al. 1984) for phaseolin (the vicilin-like protein from french bean seeds) although the number of lysines available for potential crosslinking is substantially greater in convicilin. Therefore, these results indicate that convicilin is trimeric in contrast to a previous report (Croy et al. 1980) suggesting a tetrameric structure based on empirical molecular weight measured by column chromatography. Presumably, the N-terminal extension of convicilin has not interfered with the strong trimeric associations of the vicilin-like C-terminal region.
Expression of pea convicilin in tobacco

The complete 7.5kb EcoR1 fragment containing the convicilin gene was ligated into binary vectors, either BIIN19 (Bevan 1984) or pGA492 (An 1986), to make BDL5 or pAG1, respectively. These constructs were used to transform tobacco leaf pieces using Agrobacterium-mediated infection. Plants labeled "G" were transformed with BDL5 and those labeled "H" were transformed with pAG1. The numbers after these letters refer to individual plants that arose from separate transformation events.

Convicilin protein was assayed in seed extracts by the Western blot procedure using antibodies prepared in rabbits against pea vicilins. Of eight "H" series plants examined, four (H1, H4, H6, and H7) contained both the intact gene and had detectable convicilin in their seeds (data not shown). The remaining plants contained neither a copy of the intact gene nor observable amounts of convicilin. A total of six "G" series plants were assayed and all had convicilin in their seeds. Two of the "H" series plants (H6 and H7) and four "G" series plants (G1 to G4) were selected for further work.

Genomic DNA isolated from pea seedlings and the leaves of these six putative tobacco transformants was analyzed by Southern blotting followed by hybridization with nick-translated pEN6 insert (Fig.4). Pea DNA digested with EcoR1 (Fig.4, lane 9) showed hybridization to a number of bands in agreement with those obtained by other workers using a convicilin cDNA clone (Domoney and Casey 1985; Ellis et al. 1986). A major band of 7.5kb was identical in size to the insert of p5.1 (compare Fig.4, lanes 1 and 9). Two weaker hybridizations of ~14kb and ~5.5kb were also observed. The band of ~14kb has been identified as containing a gene coding for a convicilin-related protein (Ellis et al. 1986). The band at ~5.5kb has not been characterized.

DNA isolated from the six putative transformants showed hybridization to an EcoR1 band identical in size to the major band found in pea (Fig.4, lanes 2 to 7) confirming that the complete gene had been correctly transferred. A second, weaker hybridization of ~4.6kb was
Figure 4. Southern blot of convicilin gene in pea and transgenic tobacco plants. DNA was extracted from leaves of tobacco or young pea shoots and purified. Tobacco plants with a "G" prefix were transformed with the plasmid BDL5 and those with a "H" prefix were transformed with pAG1. 10 μg of DNA was digested with EcoR1, fractionated by electrophoresis on agarose and blotted onto nylon membrane. The membrane was then probed with 32P labelled insert from pEN6. Lane 1 contains an amount of p5.1 insert equivalent to one gene copy per haploid genome. DNA in lanes 2 to 7 was from G1, G2, G3, G4, H6, and H7, respectively. DNA in lane 8 was from a plant transformed with the binary vector, pGA492 without an insert. Lane 9 contains pea shoot DNA. Numbers on the vertical axis show the size (in kilobases) of DNA markers (λ DNA digested with HindIII).
also observed in all plants and was also found in a plant that had been transformed with the binary vector pGA492 alone (Fig.4, lane 8). This band was probably caused by hybridization between the pea convicilin probe and an unidentified tobacco sequence, as it was common to plants transformed with the two different vectors.

When extracts of seeds from untransformed tobacco plants were subjected to Western blotting, a single protein of Mr~47,000 was detected (Fig.5, lane 8). When seed extracts from convicilin-transformed plants were similarly treated, a major protein identical in size to pea convicilin (Mr 70,000) was observed (compare Fig.5, lane 1 with lanes 2 to 7). A number of other polypeptides, principally of Mr~50,000, were also seen. Convicilin in pea, undergoes little if any cleavage (Chrispeels et al. 1982a), thus this processing is unique to tobacco. To investigate these polypeptides further, seed proteins from transformant H7 were analysed by two-dimensional electrophoresis before Western blotting (Fig.6). The convicilin polypeptides, with a pI of 5.6 to 5.8 (Croy et al. 1980), were well separated from the cross-reacting tobacco proteins (Fig.5, boxed) which were more basic. About twenty separate convicilin polypeptides were observed. Many of these occurred as different isoelectric forms of eight discrete molecular weights between Mr~70 and 50,000.

When the level of convicilin in the seeds of the six transformed plants was compared by densitometry, it was found that there was a twofold range in amount of accumulated protein (Fig.5, lanes 2 to 7). Plants H6, H7, G1 and G3 all have 0.5 to 1 copy of the convicilin gene (Fig.4) by comparison to pea and a reconstruction using purified p5.1 insert loaded with an amount equivalent to one copy (Fig.4, lane 1). G2 and G4 have 2 copies of the gene. Western blots of extracts of mature seeds (Fig.5) showed G2 and G4 have approximately double the amount of convicilin that was found in the other plants. Thus, there appears to be a correlation between gene copy number and level of expression.

An estimate of the amount of convicilin accumulated by plant H7 (a plant with a single copy of the gene) was obtained by comparing densitometer tracings of a concentration series of purified vicilin with a known amount of H7 total seed protein. These indicated that convicilin
Figure 5. Convicilin protein accumulated in mature seeds of pea and transgenic tobacco plants. Protein (5μg) from pea seeds and 150μg of tobacco seed protein was fractionated on an SDS-polyacrylamide gel and then electroblotted onto nitrocellulose membrane. Convicilin was detected by reacting the nitrocellulose with antiserum against total pea vicilins. Tobacco plants with the prefix "G" were transformed with the plasmid BDL5 and those labelled "H" were transformed with pAG1. Lane 1 is a pea seed extract. Lanes 2 to 7 are H6, H7, G1, G2, G3 and G4, respectively. Lane 8 contains protein from an untransformed tobacco plant (Wisconsin 38). Numbers on the vertical axis indicate Mr x10^{-3} of the polypeptides.
Figure 6. Analysis by two-dimensional separation of convicilin in the mature seed of transgenic tobacco plant H7. Protein (150μg) extracted from mature seeds was separated by isoelectric focussing then SDS-polyacrylamide gel electrophoresis (after the method of Hughes and Galau 1985). The proteins were electroblotted onto nitrocellulose and probed with anti-convicilin antiserum as described in the legend to Fig.5. Boxed polypeptides are those which are found in non-transformed tobacco seeds. Numbers on the vertical axis are polypeptide sizes (Mr x10^-3).
comprised about 2% of the total seed protein in this plant.

Developmental regulation and organ specificity

In peas, convicilin is expressed in cotyledons at later stages of seed development (Chandler et al. 1984) and has not been found in any other organ studied (Higgins et al. 1988). We tested whether the same regulatory mechanisms were functional when the gene was transferred to a new host by examining its temporal expression and organ specificity in transgenic tobacco. Seeds were harvested at specified days after flowering (DAF) and the developmental profile of convicilin accumulation was determined by Western blot analysis (Fig. 7). The vicilin-related polypeptide found in untransformed tobacco seeds was observed at the earliest harvest (Fig. 6, lane 2). This polypeptide, as has been proposed previously (Higgins et al. 1988), was synthesized as a Mr~55,000 precursor that was slowly processed to the mature Mr~47,000 form. The Mr~70,000 kd band, characteristic of convicilin, first appeared at 15 DAF. Between 15 and 19 DAF there was little further accumulation although some processing to Mr~50,000 did occur. Most of the convicilin was accumulated between 19 and 24 DAF, by which time, the level and pattern characteristic of the mature, fully dry seed was attained.

The level of convicilin mRNA throughout seed development was determined by S1 protection assay (Fig. 8). Messenger RNA was first observed at 15 DAF and reached a maximum at 17 DAF. This level was held approximately constant to at least 21 DAF, a pattern which is also true of convicilin mRNA accumulation in pea (Chandler et al. 1984). Convicilin mRNA was not detected in seeds of plants not transformed with convicilin (data not shown). The size of the DNA fragments protected from S1 nuclease was the same in both pea and tobacco indicating that the start sites of transcription used in these two species was identical. The level of mRNA detected in tobacco seeds at 17 DAF was approximately 0.5% of the level detected in pea seeds of the same age.

The organ specificity of convicilin gene expression was determined by mixing approximately
Figure 7. The pattern of accumulation of convicilin polypeptides during seed development in a transgenic tobacco plant (H7). Seeds were harvested on the specified days after flowering, protein extracts were made and 150μg of protein was fractionated on an SDS-polyacrylamide gel and then immunoblotted as described in the legend to Fig.5. Lane 1 contains protein from mature, untransformed tobacco seeds. Lanes 2 to 8 contain protein from the seeds of plant H7 at 11, 13, 15, 17, 19, 21 and 24 days after flowering. Lane 9 contains protein from mature H7 seed. The numbers on the vertical axis are polypeptide sizes (Mr x10^{-3}).
Figure 8. The level of convicilin mRNA in seeds of a transformed tobacco plant (H7) during seed development as determined by S1 protection assay. Lanes 2 to 7 contain 50μg of total RNA isolated from developing seeds at 11, 13, 15, 17, 19 and 21 DAF, respectively. The RNA was first hybridized to a single-stranded DNA probe, uniformly labelled with $^{32}$P. The probe contained 169 bases of the coding region and at least 800 bases of the 5' flanking region. The hybrids were then digested with S1 nuclease and the products fractionated on a 5% denaturing acrylamide gel before autoradiography to detect the protected fragments. Lane 1 contains $^{32}$P-labelled probe and 50μg of tRNA digested with S1. Lane 8 contains 1μg of pea RNA isolated from seeds at 17DAF. The numbers on the vertical axis refer to the sizes of the protected fragments in bases.
1 mg of protein extracted from the leaves, roots or stems of pea or tobacco transformant H7, and a lesser amount of seed extract (150 μg and 8 μg of tobacco and pea, respectively) with anti-vicilin immuno-affinity beads (Chrispeels et al. 1982b). Protein (1 mg) extracted from H7 callus was treated similarly. After washing to remove unbound protein, the beads were loaded onto an SDS-polyacrylamide gel and fractionated prior to Western blotting. Convicilin bands were only detected in the two seed extracts and not in any other organ of either pea or transgenic tobacco (data not shown).

Discussion

In this paper, we report the structure and deduced amino acid sequence of pea convicilin as well as partial sequences of both genomic and cDNA clones encoding convicilin and the expression of the convicilin gene in transformed tobacco plants. Convicilin was shown to be a trimeric protein with a low α-helical content. Both of these physical characters are typical of 7S globulins (Blagrove et al. 1984). An earlier report indicating that convicilin was a tetramer (Croy et al. 1980) was based on the relatively low resolving power of gel filtration chromatography. In the present study, both sedimentation equilibrium centrifugation and chemical crosslinking indicate that convicilin has a trimeric structure. The close serological relationship between vicilin and convicilin (Croy et al. 1980; Spencer et al. 1983) is explained by the fact that convicilin represents vicilin with an N-terminal extension (Casey et al. 1984; Bown et al. 1988). Such terminally extended proteins are also found in other seeds, such as the α' subunit of β-conglycinin of soya bean (Doyle et al. 1986) and the α-globulins of cotton (Chian et al. 1987).

Although the genomic and cDNA sequences reported here are not identical (4 differences in 576 common nucleotides), we suggest, for reasons given below, that this variation arises from the presence in our line of peas of at least two alleles for this gene rather than a family of closely related genes. The flanking regions of the convicilin gene show the features common to RNA polymerase II transcription units from plants and animals such as the conserved "CAAT" and "TATA" boxes upstream of the coding region and repeated polyadenylation
signals 3' to the end of translation. A single intron of 151 bp was found in the partially-
sequenced genomic clone at the corresponding position, relative to the deduced amino acid
sequence, as intron one of a Mr 50,000 vicilin gene (Higgins et al. 1988). Typically, 7S
storage protein genes are interrupted by 5 introns located in similar places in the sequence
(Doyle et al. 1986). Another feature found in the 5' flanking regions that is common to this
class of genes is the "vicilin box" (Gatehouse et al. 1986; Bown et al. 1988; Higgins et al.
1988). This sequence of about 40 bp, starting at position -137 in the convicilin gene with the
hexanucleotide GCCACC, has been suggested as determining tissue specificity of expression
in this gene family (Gatehouse et al. 1986). Although this sequence element may have a role
in regulating expression of this gene family, it failed to bring about the expression in transgenic
plants of a reporter gene (Chloramphenicol Acetyl Transferase) when used in a construction to
test its ability to act as a seed promoter. Additional sequences are required for activity (E.
Newbigin, unpublished observation).

Two genes have now been identified in pea that hybridize to convicilin cDNAs (Bown et al.
1988; this paper). On Southern blots of genomic DNA digested with EcoR1, these genes
occur as bands of ~14kb and 7.5 to 9kb (depending on variety)(Domoney and Casey 1985;
Ellis et al. 1986) with the lower band hybridizing more strongly than the upper band. These
two bands segregate together in crosses and are therefore linked (Ellis et al. 1986; Casey et
al. 1986). A comparison between the sequence presented here and that given by Bown et al.
(1988) for the gene on the ~14kb EcoR1 fragment shows that the two sequences are 95%
identical after gaps are introduced to increase similarity (data not shown). The largest
difference between these two genes results from the fact that the sequences duplicated in the
first exon of the gene on p5.1 are found only once in the sequence of Bown et al. (1988). This
difference in the length of exon 1 accounts for the discrepancy in molecular weight predicted
for the mature polypeptides (68.2kd here and 63.9kd in Bown et al. 1988). In their paper,
these authors tentatively ascribe the gene contained on the ~14kb fragment to the Cvc locus.
The Cvc locus (Matta and Gatehouse 1982; presumably identical to the Vca locus defined by
Thomson and Schroeder 1978) is the only locus so far defined for convicilin and was identified
on the basis of variation in the mobility of the Mr 70,000 polypeptides (Mr 75,000 in Thomson
and Schroeder 1978) on SDS-polyacrylamide gels. Several lines of evidence suggest that the
gene described by Bown et al. (1988) is unlikely to be the gene at the Cvc locus. The gene on
p5.1 is at least 98% identical with the two cDNAs so far published that encode convicilin
(pCD59 in Casey et al. 1984; pPS15-28 in this paper). Overall, the gene reported by Bown et
al. (1988) and pCD59 are 94% identical. When p5.1 was integrated into the tobacco genome
it gave rise to a protein of identical size to convicilin in pea (Fig.5). The gene described by
Bown et al. (1988) encodes a protein that is smaller (63.9kd) than either the reported size of
convicilin (Mr~70,000; Casey and Sanger 1980; Croy et al. 1980) or the protein encoded on
p5.1 (68.2 kd). On these bases, we propose that the gene found on p5.1 is identical to the
Cvc locus and that the gene described by Bown et al. (1988) is a linked gene coding for an as
yet undescribed, convicilin-related protein. Similarly, since pPS15-28 and p5.1 resemble each
other closely (>99% identical) it is probable that pPS15-28 represents the product of another
allele of the Cvc locus rather than that of a distinct gene.

When integrated into the tobacco genome, the gene on p5.1 showed all the characteristics
that are typical of its expression in pea. Its expression was confined to the seed and it was
only detected during the later stages of cotyledon development (Fig.7). Although the mRNA
was present at an approximately constant level between 15 and 21 DAF as in pea (Chandler et
al. 1984), convicilin protein was preferentially accumulated somewhat later in development.
The same transcription start points were used in both pea and tobacco (Fig.8). A similar level
of fidelity of expression has been shown for a number of other 7S storage protein genes when
transferred to new host species (Sengupta-Gopalan et al. 1985; Beachy et al. 1985; Bray et al.
1987; Higgins et al. 1988). However, processing of the protein once expressed is, in many
cases, different from that in the original host. For example, when the β subunit of phaseolin
(Sengupta-Gopalan et al. 1985) or the α' subunit of β conglycinin (Beachy et al. 1985) were
expressed in transgenic hosts (either tobacco or petunia), as much as 50% of the protein
found in the mature seed that was derived from the introduced gene was smaller than the
expected size. This was also true of convicilin which was partially processed in tobacco seeds
from a protein of identical size to that found in pea seeds to polypeptides of lower molecular
weight (principally Mr~50,000). Based on two-dimensional Western blots, as many as twenty
separate isoelectric variants of eight different molecular weights were detected. These differences in isoelectric point may reflect separate post-translational modifications or may arise as an artefact of sample preparation and/or analysis. However, it is clear that they must all arise from a single primary translation product.

Convicilin comprised about 2% of the total protein present in mature seeds of transgenic plants such as H7 that contained a single copy of the convicilin gene. This level is similar to that found by Beachy et al. (1985) in transgenic experiments with α',β-conglycinin, a well-expressed gene from soybean. When a gene encoding a Mr 50,000 vicilin (pEN2) was transferred to tobacco (Higgins et al. 1988), the best expressing plant (B3), which contained about 30 copies of the gene, had accumulated vicilin to about 0.5% of the total seed protein by maturity. The difference in the level of accumulation in transgenic plants of the two types of vicilin may reflect dissimilarities in the transcriptional strength of their promoters. The 68.2 kd protein, convicilin, is encoded by a single gene and accounts for 10-20% of the total vicilin in a mature pea seed. The remainder, made up of the Mr 50,000 vicilins, is encoded by a multigene family of approximately 18 members. It is possible that the convicilin genomic clone, p5.1, has a sequence or sequences that the vicilin gene used in the earlier study (Higgins et al. 1988) lacks which results in a higher level of expression of the convicilin gene compared to the vicilin gene. Further work will be required to identify the DNA sequences responsible for regulating the level of expression of these genes.
Chapter 4

Expression of vicilin promoter:cat chimeric genes in the seeds of transgenic tobacco plants.
Introduction

The 5' flanking regions of plant genes contain short (<100 bp) regulatory elements that are required for tissue-specific expression (for a review, see Kuhlemeier et al. 1987). Frequently elements with the same sequence have been found in the upstream regions of functionally-related genes such as those responsive to heat-shock (Schoffl et al. 1986) or light (Simpson et al. 1986; Fluhr et al. 1986). Short sequences, conserved among members of closely-related storage protein genes families, have been identified in a number of cases. Legumin-type (11S) genes share a sequence of 28 bases located around -100 relative to the start of transcription called the legumin box (Baumlein et al. 1986). Cereal genes coding for polypeptides of the prolamine (alcohol-soluble) class contain a sequence called the -300 box (Forde et al. 1985) and a region called the vicilin box has been described for genes encoding proteins of the 7S class (Gatehouse et al. 1986; chapter 2). The sequence at the 5' end of the vicilin box, GCCACCTC, is similar to the E1A enhancer core-sequence from Adenovirus (Lycett et al. 1984) and is also found in the upstream regions of legumin genes (although it is not associated with the legumin box) and the seed lectin genes of french bean (Voelker et al. 1987). In each case, the suggestion has been made that the conserved sequences are in some way involved in the expression or regulation of the genes to which they are attached.

Plant transformation provides a good system with which to assay the function of regulatory elements in vivo. Putative regulatory elements can be isolated from their normal coding regions and identified by deleting or modifying candidate sequences and introducing them, together with an adjacent reporter gene, into a plant host and assaying for activity (Timko et al. 1985; Simpson et al. 1986).

In this chapter, the question of whether the conserved sequences in the 5' flanking region of a pea vicilin gene can mediate the expression of a linked coding region was addressed. The results showed that these conserved sequences, although they may be necessary, were not sufficient and that sequences upstream of position -227 were required for seed-specific expression.
Methods

Construction of the vicilin promoter:cat chimeric genes

A number of constructions were made using fragments from the upstream region of the vicilin gene pEN2. All DNA manipulations were done according to the standard methods outlined in Maniatis et al. (1982). The constructions were based on the plasmid pΔ35SCN which was a generous gift of Dr J.C. Walker. Details of the way in which pΔ35SCN was made are given in Walker et al. (1987). pΔ35SCN contains 45 nucleotides upstream of the transcription start site of the CaMV 35S promoter (bases 7391 to 7436, numbering as in Frank et al. 1980) linked to the cat gene coding sequence (including 27 bp of the 5' untranslated leader) from Tn9 and the 3' transcription termination sequence from nopaline synthase (nos) (Bevan et al. 1983). A fragment (Δ373), containing the sequence between -49 and -227 of a vicilin gene (pEN2), was isolated from a library of randomly-generated clones used in sequencing pEN2 (see appendix 1). This fragment included all the regions found to be conserved among a number of genes for 7S proteins except the TATA box and transcription start site (see chapter 2). Δ373 was cloned upstream of the cat gene in both orientations to give intermediate clones pEN4 and pEN5 (see Fig.1a). The orientation of these insertions was mapped with restriction enzymes and by sequencing a BamH1 fragment from pEN4 (Fig.1b). Δ373 was inserted in the correct orientation, with respect to the direction of transcription, in pEN4 and in the reverse orientation in pEN5 (Fig.1a). These chimeric gene fusions were cloned into the binary vector, pGA482 (An 1986) to give pEN12 and pEN14 (Fig.1a). The inactive cat gene from pΔ35SCN was cloned into pGA482 to give the promoterless control, pEN13 (Fig. 1a). These plasmids were transferred to Agrobacterium tumefaciens by conjugation and the recombinants analysed by Southern blotting (Fig.1a).

The complete 5' flanking sequence from the vicilin gene (pEN2) was cloned upstream of the cat gene in pEN4. This was done by replacing the sequence upstream of the Nde1 site in pEN4 with an Nde1 fragment from pEN2 (see Fig.1c). The intermediate plasmid was called pEN8. The chimeric cat gene from pEN8 was cloned into pGA482 to give pEN18 (Fig.1c).
Figure 1a (two pages). Construction of the *Agrobacterium* binary vectors pEN12, pEN13 and pEN14. pEN12 and pEN14 contain vicilin promoter fragment:cat chimeric genes and pEN13 includes the promoterless cat control. □ represents the fragment of the vicilin upstream region (Δ373) which contains the sequence between -49 and -227 with ▼ and ▼ showing the approximate locations of the vicilin and CAAT boxes, respectively. The plasmid pΔ35SCN (Walker et al. 1987) carries the cat gene (□) together with the transcription termination region from the nopaline synthase gene (nos)(▼) and the sequence between +1 and -45 of the CaMV 35S promoter (◆). □ shows the approximate position of the 35S TATA box. The horizontal lines represent polylinker sequence with important restriction enzyme sites shown by vertical lines. The HindIII-EcoR1 fragment, Δ373 was cut with HindIII and the EcoR1 site was removed by treatment with the large fragment of DNA polymerase I (Klenow's fragment). The resulting DNA was ligated into HindIII-cut pΔ35SCN to give intermediate plasmids, pEN4 and pEN5. The orientation of the inserted Δ373 fragment is shown by the arrows. The Sma1 to HindIII inserts of pEN4, pEN5 and pΔ35SCN were ligated into HindIII and Sca1- cut pGA482 (An 1986) to give pEN12, pEN14 and pEN13, respectively. The expected size (in kilobases) of the DNA fragments resulting from digestion of these constructions with BamH1 are given in parentheses. B\(\text{l}\) and B\(\text{r}\) refer to the left and right T-DNA borders; T\(\text{c}\)\(\text{R}\), the bacterial tetracycline resistance gene; the plant kanamycin resistance marker constructed from the nos control regions and neomycin phosphotransferase coding region is indicated by the closed box.
Vicilin (pEN2) fragment

Δ373

HindIII
BamHI
NdeI
EcoR1
-227
-49

HindIII-cut and blunt-end with Klenow's Fragment

BamHI
HindIII
SmaI
EcoR1
-45
+1

Open plasmid with HindIII and phosphatase

HindIII

---

pEN4

BamHI
SmaI
EcoR1
-227
-49

HindIII

---

pEN5

BamHI
SmaI
NdeI
EcoR1
-49
-227

HindIII

---

Cut with SmaI and HindIII and ligate into HindIII and Scal-cut pGA482

---

pEN12, 14 (pGA482)

B1
B2

(pGA482)

B1
B2

---

pEN13 (pGA482)

B1
B2

---

(3.9)

(9.2)

(3.9)

(9.2)

$T_c^R$
Figure 1a (second page). Southern blot analysis of pEN12, pEN13 and pEN14. DNA, extracted from Agrobacterium transformed with these constructions, was restricted with BamH1, the fragments separated on a 0.8% agarose gel and blotted onto nitrocellulose. These filters were hybridized with 32P-labelled Δ373 and pΔ35SCN (CAT) inserts. The numbers on the vertical axis are the size (in kilobases x10^-3) of standard DNA fragments (λ digested with HindIII).
Figure 1b. Sequence of the small BamH1 fragment from pEN4 containing the vicilin promoter fragment A373 (in bold type) and the sequence between +1 and -45 of the 35S promoter (in italics). Underlined nucleotides result from the ligation of these two fragments. Numbers above and below the sequence give the position relative to the start of transcription in the vicilin and 35S promoters, respectively.
GAAGACAATA ATAAAGCATC CTCCTTTTCC ATAAAGATGT CCAAATTCAT

CAAATTCAAA CAAAACTCCA CCACCCAAAGT AATCTTTCTT TCATTCTGCC

ACTTCAATT TGTACATT TT AACACACGTC CATATGATG GCACAACATG

GCCAACTGTGTT GGTGACATGT AATTATAG GGGAAATTGAC AAGACCCCTTC

CTCTATATAA GGAAGTTCAT TTCAATGGGA GAGAAC
Figure 1c (two pages). Construction of the *Agrobacterium* binary vector, pEN18. pEN18 contains the sequence present in the vicilin gene upstream of position -49, linked to the *cat* gene in pΔ35SCN (see Fig.1a). The upstream region of vicilin (□) was isolated by digesting the plasmid containing the vicilin gene (pEN2) at the Nde1 site within the sequence and at a second site within the vector, pUC8 (----). The intermediate plasmid, pEN4, (see Fig.1a) that included the vicilin promoter sequence from -49 to -227 was also cut at the two Nde1 sites and the large fragment containing the *cat* coding region was ligated to the vicilin upstream region from pEN2. The resulting plasmid was called pEN8. The HindIII insert of this construction was ligated into HindIII-cut pGA482 and named pEN18. The numbers in parentheses indicate the expected size of fragments created by digestion with HindIII (numbers within the circle) or BamH1 (numbers outside the circle). □ represents the coding region of the vicilin gene. Other symbols are as given in the legend to figure 1a.
Cut pEN2 and pEN4 plasmids with Nde1 and ligate the upstream region fragment from pEN2 to the coding region fragment of pEN4.

Cut pEN8 with HindIII and ligate fragment into HindIII-cut pGA482.
Figure 1c (second page). Southern blot analysis of pEN18. DNA, extracted from *Agrobacterium* transformed with pEN18, was restricted with either HindIII or BamH1, fractionated on a 0.8% agarose gel and transferred to nitrocellulose. These filters were hybridized with $^{32}$P-labelled Δ373 or pΔ35SCN (CAT) inserts. The numbers on the vertical axes are the sizes of standard DNA fragments (λ digested with HindIII). The band observed between the 9.4 and 23.1 kb size markers in the HindIII lane probed with CAT results from hybridization between the *nos* termination signal on the probe and that on the plant selectable marker (neomycin phosphotransferase).
This construction was used as a positive control for seed-specific expression.

**Transformation of tobacco plants**

The constructions outlined in the results section were cloned into the binary vector, pGA482, (An 1986) and mobilized into the *Agrobacterium* receptor strain, LBA4404, as described (Ditta *et al.* 1980). Transformation and regeneration of tobacco plants was done according to the procedures outlined in appendix 1.

**Assays for CAT activity in transgenic tobacco seeds**

Flowers of transgenic tobacco plants were tagged at anthesis. Capsules were harvested at specified days after flowering and CAT was assayed as described by Herrera-Estrella *et al.* (1983). Two hundred seeds were collected from each harvested capsule and ground in 400μl of 0.25M Tris-HCl, pH 7.5. This was incubated at 60°C for 10 minutes to reduce background activity (Colot *et al.* 1987). The homogenate was centrifuged and the protein concentration of the supernatant determined by the dye-binding method of Bradford (1976). Reactions were carried out in 150μl with 0.04μCi of [14C] chloramphenicol (Boehringer, Mannheim), 0.8mM acetyl coenzyme A and (except where otherwise stated) 40μl of the seed extract for 45 minutes at 37°C. The acetylated products of the reaction were separated from the non-acetylated chloramphenicol by thin-layer chromatography (TLC) on a silica plate in methanol:chloroform (5:95,v/v). The thin-layer plate was enhanced for fluorography by spraying with 2-methyl naphthalene containing 0.4% 2,5 diphenyloxazole (PPO), dried, and exposed to X-ray film (Fuji film RX, Fuji Photo Film Co, Tokyo, Japan) overnight at -80°C.

**DNA isolation**

DNA was isolated from *Agrobacterium tumefaciens* by the method of Dhaese *et al.* (1979). Southern blots were done as described (Southern 1975). DNA was isolated from the leaves of transgenic tobacco plants as described by Sutton (1974) and Southern blotting was done as outlined in chapter 3.
Results

Assay of transgenic plants for presence of chimeric genes

Tobacco plants transformed with the cat chimeric genes were regenerated from tissue culture. In addition, a plant containing a chimeric cat coding region with a 35S promoter and nos 3' transcription stop signal was obtained from Dr D. Llewellyn. This plant expressed cat in all organs and was used as a positive control for constitutive expression. A summary of the cat-chimeric constructions used is given in figure 2. Results from plants transformed with these constructions were compared with those obtained from a plant transformed with the binary vector pGA482 alone.

DNA was extracted from the leaves of transgenic plants, purified and analysed by Southern blotting after restriction endonuclease digestion with either BamH1 (pEN12, pEN13, pEN14 and 35SC transformed plants) or HindIII (pEN18 transformed plants). The Southern blots were hybridized with a radioactive DNA probe containing the cat coding region and nos transcription termination signal. The results of these experiments are shown in figure 3. No hybridization was observed between the probe and DNA isolated from an untransformed Wisconsin 38 tobacco plant. Two main bands were found with DNA from a plant transformed with the vector alone. This was caused by hybridization between the nos 3' sequence of the probe and that present on the vector and indicated that the nos sequence had integrated at least twice into the genome of this plant.

A single band of approximately 3.9 kb was seen with DNA isolated from two plants (13/1 and 13/2) transformed with the promoterless cat construction, pEN13. This was the size expected for the BamH1 fragment based on the restriction map of this plasmid (see Fig.1a) and showed that the construction had been transferred without detectable re-arrangement to these plants.

DNA from a plant harbouring the construction 35SC showed hybridization to three bands of approximately 7, 3 and 1 kb in size. These bands were those expected for this construction based on previous plasmid mapping (D. Llewellyn, personal communication).
Figure 2. Constructions used for seed expression studies in transgenic tobacco plants. The chimeric genes contain different lengths of either the CaMV 35S ( □ ) or vicilin ( □ ) upstream regions linked to the cat coding region and 3’ regulatory signals of the nos gene ( S, ■ ). Numbers indicate the position of nucleotides relative to the start of transcription of the vicilin (pEN2) gene.
Figure 3 (two pages). Southern blots of genomic DNA from plants transformed with the constructions shown in figure 2. DNA was digested with BamH1 (HindIII in the case of plants transformed with the construction pEN18) and 10μg of DNA was electrophoresed on a 1% agarose gel. Lambda DNA cut with HindIII was used as a size marker. The fragments were transferred onto nylon membrane as outlined in the Methods section and the filters probed with $^{32}$P-labelled insert from pΔ35SCN (see Fig.1a). Numbers under the figure refer to the individual transformants, thus 13/1 indicates that this plant was plant number one from the series of plants transformed with pEN13.
Four plants, transformed with the construction pEN18, were examined by Southern blotting (Fig.3). A band of 4 kb, the size expected for this cat chimeric construction with the full length vicilin promoter (Fig.1c), was observed in all plants and indicated that the construction had been successfully transferred. However, there were considerable plant-to-plant differences in the intensity of these bands, showing that these plants had not integrated the same number of copies of this construction into their genomes. Although the precise number of copies was not determined, plant 18/6 contained the least number of copies (possibly only one or two) of any of the plants from this transformation series.

Seven plants were examined from both the pEN12 and pEN14 transformation series. These were the major experimental constructions and had a fragment from the 5' flanking region of vicilin that was thought to contain the necessary signals for seed expression placed in one of two possible orientations upstream of the cat coding region (see Fig.2). Southern blot analysis, using the cat-nos probe discussed previously, was expected to show a BamHI fragment of 3.9 kb in the DNA of plants that had correctly integrated these constructions. Only two plants (12/2 and 12/5) among those transformed with pEN12 showed the predicted hybridization to the probe, although faint bands of the correct size were observed on the original autoradiograph for plants 12/6 and 12/7 (Fig.3). As these plants were all kanamycin-resistant and therefore had incorporated at least the selectable marker gene from pGA482, hybridization was expected between the nos 3' end of the marker gene and that on the DNA used to probe this Southern blot (for example, refer to the pGA482 lane on Fig.3). Since the expected hybridization was not observed, it remained a possibility that other plants of this series also contained this construction. As was seen with the pEN18 series, a different number of copies was integrated by each of the plants with 12/2 having more copies than the others.

Of the seven plants transformed with pEN14, all except one (14/1) appeared to have integrated the construction correctly (Fig.3). Plant 14/1 appeared to have a re-arrangement of the construction as neither of the observed bands was the expected size. The more intense
Figure 4. Comparison of CAT activity during seed development for a selection of transgenic tobacco plants. Seeds from tobacco plants transformed with one of the constructions shown in figure 2 or the binary vector pGA482 alone were harvested on the day following flowering specified by the numbers under the figures. CAT activity at each time point was assayed on a volume of extract equivalent to 20 seeds. CAM and Ac refer to the non-acetylated and acetylated forms of chloramphenicol.
assays reported here used equal amounts of tissue whereas the developmentally-regulated accumulation of vicilin was based on equal amounts of protein.

No CAT activity was detected in any of the harvests of seed from plants 12/5 or 14/8 which contained the vicilin upstream region:cat-chimeric constructions, pEN12 and pEN14, respectively (see Fig.4). These experiments indicated that the conserved sequences from the 5' flanking region of vicilin were insufficient to direct the transcription of a linked gene although other transformed plants of this series would have to be examined to exclude the possibility that the results were brought about by differences in the site at which the transferred genes were inserted into the host genome.

Capsules were harvested at 17 to 18 DAF from the remaining plants, that is those analysed by Southern blot in figure 3 but not selected subsequently for the experiments reported in figure 4. This time period was chosen on the basis of the results obtained with plant 18/6 (see Fig.4). Extracts were made from the seeds as previously described and the level of CAT in these extracts was determined. The results of these experiments is shown in figure 5. Of the seed extracts examined, only those from plants transformed with pEN18 (18/2a,18/2b and 18/5) showed levels of CAT that were higher than those of the control (13/2). An exception to this was plant 14/4. This plant previously was shown to have more copies of the construction pEN14 integrated into its genome than any of the other plants from this transformation series (see Fig.3). It is unlikely that the level of CAT activity seen in extracts from this plant was due to cat expression driven by the vicilin upstream region present in the reversed orientation in this construction for reasons to be discussed below. Possibly this activity was caused by an accumulated low level of transcription from the 35S TATA box and reflected the larger number of copies of the chimeric gene present in this genome.

There was little variation found between plants transformed with pEN18 in the levels of CAT present in the seeds (compare Fig.4, 18 DAF lane of plant 18/6 with Fig.5) although there were differences in the gene copy number (see Fig.3). Previously it had been found that the amount of vicilin produced by the seeds of plants transformed with pEN2 (the vicilin gene) was related
Figure 5. Comparison of CAT activity found in the seeds of transgenic tobacco plants harvested at 17-18 DAF. Each assay was done on a volume of extract equivalent to 20 seeds. Numbers under the figure refer to individual transformed plants.
Figure 6. Comparison of the level of CAT activity found at 18 DAF in the seeds of transgenic plants 18/6 and 13/1. 18/6 was transformed with construction pEN18 that contained the full vicilin upstream region attached to the cat gene and 13/1 was transformed with the promoterless control, pEN13. 200 seeds were harvested from each plant and ground in 400μl of extraction buffer. The level of CAT in the indicated amounts of this extract was determined.
to the number of copies integrated into the genome (see appendix 1). Even though the number of plants examined here was small (four in total), there appeared to be no relationship between copy number and level of CAT.

Although none of the plants transformed with pEN12 or pEN14 showed levels of CAT in their seeds that were above those in the control plants transformed with either pGA482 or the promoterless cat construction, pEN13, it was possible that these plants expressed cat to an extent below that which could be resolved by this assay. To determine what level of CAT activity, resulting from transcription directed by the vicilin promoter, could be distinguished from background, decreasing amounts of an extract made from seeds harvested at 18 DAF from plant 18/6 were used in a series of CAT assays. The results of this experiment are shown in figure 6. The level of CAT activity assayed in 5μl of this extract was above that in 40μl of an extract made from 18 DAF seeds from plant 13/1. This indicated that an activity down to approximately 12% of that resulting from the expression of the full length vicilin upstream region could be detected. It could be concluded that if constructions pEN12 and pEN14 were able to direct expression of the linked cat gene, they did so at a level that was less than 12% of that directed by pEN18.

Discussion

This chapter describes experiments set up to discover whether the vicilin box (bases -90 to -131 of pEN2; see chapter 2) and other conserved sequences found in the upstream regions of 7S storage proteins were sufficient to bring about seed-specific expression of an adjacent coding region. This was done by testing in transgenic tobacco a number of constructions that differed in the length or orientation of the 5' flanking sequence used from pEN2.

Two constructions were made to test if the fragment, Δ373, which contained sequences between -49 and -227 of pEN2 could direct the expression of a linked coding region (cat). This region was similar to the one used by Chen et al. (1988; bases -78 to -257 from an α', β-conglycinin gene) to demonstrate the existence of a seed-specific enhancer. The Δ373
fragment contained all the regions of sequence conserved between 7S genes upstream of the TATA box (see chapter 2). This was ligated in both orientations in front of a construction (pΔ35SCN) containing the TATA box and transcription start site from the 35S promoter of CaMV, the cat coding sequence from Tn9 and the Nos 3' termination region to make pEN12 and pEN14. PEN12 contained the A373 sequence in the same orientation as originally found in the 5' flanking region of pEN2 and pEN14 had this sequence in the opposite orientation. It seemed unlikely that pEN14 would actively express in seeds given that the CAAT box had been placed at a greater distance from the TATA box and that its orientation had been reversed. Large alterations to the spacing between the TATA box and various upstream elements have been shown to be deleterious to expression (McKnight 1982; Takahashi et al. 1986). A further construction pEN18 was made in which the pEN2 sequences upstream of position -227 were added to those already in pEN12. Of these constructions, only plants transformed with pEN18 showed levels of CAT that were above background (figure 8) indicating that sequences upstream of -227 were required for expression. Plants transformed with pEN12 or pEN14, if they expressed at all, were expressing CAT at levels less than 12% of the level found in pEN18 plants. This suggests that the organization of the promoter region in vicilin is different to that described for α', β-conglycinin (Chen et al. 1986). They reported that sequences to -257 were all that were required for a normal level of expression (as given by a construction with 8.5 kb of 5' sequence) and that sequences to -208 gave expression at 69% of this level. Expression levels in this gene were correlated with the number of copies of the sequence "AG/ACCCA" present in each of the deletions. This hexanucleotide does not occur on either strand of the upstream regions of pEN2 (but was found twice in a search of the upstream regions of pea convicilin, a similar protein to α', β-conglycinin) so cannot play a role in the expression of this gene.

This chapter has shown that the vicilin box was not sufficient (together with the TATA and CAAT boxes) to express a gene in a seed-specific manner. This does not indicate, however, that the sequence is not required for organ-specific regulation. As demonstrated in the study by Stougaard et al. (1987) on the promoter for the lbc3 gene (a gene for leghaemoglobin from soya bean) conserved sequences, found in the 5' regions of all root-nodule-specific genes,
were not adequate for expression of a linked gene and needed positive control elements to drive transcription. The elements found in the \(lbc_3\) 5' region were not absolutely required and could be replaced by the constitutive CaMV 35S enhancer. Expression of the resulting constructions (CaMV 35S and consensus sequence from \(lbc_3\) promoter) was only observed in nodules of transgenic plants and indicated that the conserved elements conferred organ specificity on the constitutive 35S enhancer. This separation between regulatory and positive control elements also has been reported for the chalcone synthase \(chs\) promoter (Kaulen et al. 1986) and the chlorophyll a/b binding protein gene \(ihcp\) promoter (Simpson et al. 1985). It is possible that this situation also applies to the pEN2 (vicilin) gene promoter as sequences upstream of -227 are required for expression. Further experiments will be required to demonstrate this. These experiments could include the linking of the CaMV 35S enhancer or a similar constitutive enhancer element upstream of the vicilin 5' region in construction pEN12.

If expression of this construction was limited to the seed, it would indicate that sequences within the -49 to -227 fragment of pEN2 act as seed-specific regulators. Furthermore, 5' end deletions or linker scanning mutations (McKnight 1982) could be used to better define the upstream elements found in the pEN2 promoter. The 5' flanking region from pEN18, which has been shown to direct seed expression, could most easily be used for these manipulations as it can be isolated as a BamH1 fragment of approximately 2 kb.

In chapter 2, I began by outlining the sequences found to be conserved among four members of the 7S storage protein genes. I assumed that sequences that were common to them would be used by the plant to bring about their appropriate expression. As these four genes had all been used in transgenic experiments it might also be assumed that these same signals would be used by the transgenic host to direct expression. This assumption has proved to be incorrect and the question is therefore asked: "If the sequences common to these genes are not the sequences recognized by the transgenic host, what are the signals in the non-conserved sequences that are recognized?". Table 1 gives a list of genes for seed-storage proteins that have been used in transgenic experiments. Only those constructions using the original control sequences and disarmed vectors have been considered. All of the gene donor species are evolutionary well-separated from the transgenic hosts. It has been estimated that
Table 1. List of seed storage protein genes that have been expressed in transgenic hosts. Only those genes that have been expressed in non-tumorogenic tissues using the original transcriptional control signal have been considered. The words "not determined" indicate that the sub-organ distribution (i.e. embryonic or endospermic) of the protein or mRNA was not examined. Unless otherwise stated, all genes were specifically expressed in the seeds of transgenic plants.
<table>
<thead>
<tr>
<th>Donor species</th>
<th>Gene name</th>
<th>Expression localized to</th>
<th>Transferred to</th>
<th>Localized in</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>phytohemagglutinin</td>
<td>(as above)</td>
<td>tobacco</td>
<td>(as above)</td>
<td>Voelker et al. 1987, Sturm et al. 1988</td>
</tr>
<tr>
<td>Soya bean</td>
<td>α', β-conglycinin</td>
<td>(as above)</td>
<td>petunia</td>
<td>not determined</td>
<td>Beachy et al. 1985</td>
</tr>
<tr>
<td></td>
<td>β, γ-conglycinin</td>
<td>(as above)</td>
<td>petunia</td>
<td>cotyledon parenchyma cells</td>
<td>Bray et al. 1987, Barker et al. 1988</td>
</tr>
<tr>
<td></td>
<td>lectin</td>
<td>seed and root</td>
<td>tobacco</td>
<td>seed and root</td>
<td>Okamuro et al. 1986</td>
</tr>
<tr>
<td>Pea</td>
<td>vicilin</td>
<td>cotyledon parenchyma cells</td>
<td>tobacco</td>
<td>mostly cotyledon (some endosperm)</td>
<td>appendix 1</td>
</tr>
<tr>
<td></td>
<td>convicilin</td>
<td>(as above)</td>
<td>tobacco</td>
<td>not determined</td>
<td>Chapter 3</td>
</tr>
<tr>
<td></td>
<td>legumin</td>
<td>(as above)</td>
<td>tobacco</td>
<td>not determined</td>
<td>Ellis et al. 1988, W. Rerie pers. comm.</td>
</tr>
<tr>
<td>Wheat</td>
<td>Low molecular</td>
<td>endosperm</td>
<td>tobacco</td>
<td>endosperm</td>
<td>Colot et al. 1987</td>
</tr>
<tr>
<td></td>
<td>and high molecular weight glutenins</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Barley</td>
<td>B-hordein</td>
<td>(as above)</td>
<td>tobacco</td>
<td>(as above)</td>
<td>Marris et al. 1988</td>
</tr>
<tr>
<td>Maize</td>
<td>23 kd zein</td>
<td>(as above)</td>
<td>tobacco</td>
<td>(as above)</td>
<td>Scherthaner et al. 1988</td>
</tr>
<tr>
<td></td>
<td>19 kd zein</td>
<td>(as above)</td>
<td>petunia</td>
<td>all plant organs</td>
<td>Ueng et al. 1988</td>
</tr>
</tbody>
</table>
tobacco and french bean are separated by about 10 million years of evolution and that monocots diverged from dicots 100 to 150 million years ago (Boulter et al. 1972; Smith 1976). It is therefore striking that the DNA-encoded signals for seed expression are correctly recognized by the transgenic host. Moreover, these genes are only expressed in embryonic or endospermic tissue depending on whether they were derived from dicot or monocot species. This contrasts with normal tobacco seed proteins which are synthesized in both embryo and endosperm. The exception to this is a 19 kd zein gene which was transcribed at low levels (but not apparently translated) in all organs of the transgenic plants examined (Ueng et al. 1988). Nor is it typical that monocot genes are expressed in dicotyledonous hosts as tobacco failed to express the gene for wheat small subunit of ribulose-1,5-bisphosphate carboxylase (Keith and Chua 1986) and maize alcohol dehydrogenase (Adh1; Llewellyn et al. 1985).

Although the storage proteins listed in table 1 are functionally related, there are few similarities in their 5’ flanking regions apart from the TATA box. One short sequence, called the RY repeat, has been found in the upstream regions of most seed-expressed genes from legumes (Dickinson et al. 1988) and some cereal protein genes (Forde et al. 1985). RY repeats are characterized by a succession of alternating purine and pyrimidine nucleotides and include the common sequence CATGCATG (Vodkin et al. 1983; Hoffman 1984). This sequence was found at position -96 in the pEN2 gene and was hence present in the pEN12 construction. As this construction was apparently inactive, the RY repeats cannot be solely responsible for the expression in transgenic plants of the genes listed in table 1.

The way in which transgenic plants interpret the idiosyncratic signals found in this wide range of genes is mysterious. Resolution of this mystery will provide information central to the understanding of transcriptional control of gene expression.
References


Craig, S., Millerd, A., Goodchild, D.J. (1980b) Structural aspects of protein accumulation in developing pea cotyledons. III. Immunocytochemical localization of legumin and vicilin using antibodies shown to be specific by the enzyme-linked immunoabsorbent assay (ELISA). Aust.J.Plant Physiol. 7,339-351.


Craig, S., Goodchild, D.J. (1984b) Golgi-mediated vicilin accumulation in pea cotyledon cells is redirected by monensin and nigericin. Protoplasma 122,91-97.


Keith, B., Chua, N.-H. (1986) Monocot and dicot pre-mRNAs are processed with different efficiencies in transgenic tobacco. EMBO J. 5,2419-2425.


Vodkin, L.O., Rhodes, P.R., Goldberg, R.B. (1983) cA lectin gene insertion has the structural features of a transposable element. Cell 34,1023-1031.

Voelker, T., Sturm, A., Chrispeels, M.J. (1987) Differences in expression between two seed lectin alleles obtained from normal and lectin-deficient beans are maintained in transgenic tobacco. EMBO J. 6,3571-3577.


Appendix 1

The sequence of a pea vicilin gene and its expression in transgenic tobacco plants

Thomas J.V. Higgins, Edward J. Newbigin1, Donald Spencer*, Danny J. Llewellyn and Stuart Craig

CSIRO, Division of Plant Industry, GPO Box 1600, Canberra 2601, Australia.
1: Department of Botany, Australian National University, GPO Box 4, Canberra, 2601 Australia.

Key words: Vicilin gene, transgenic tobacco, Agrobacterium tumefaciens
Abstract

A 5.5 kb EcoRI fragment containing a vicilin gene was selected from a *Pisum sativum* genomic library and the protein-coding region and adjacent 5' and 3' regions were sequenced. A DNA construction comprising this 5.5 kb fragment together with a gene for neomycin phosphotransferase II was stably introduced into tobacco using an *Agrobacterium tumefaciens* binary vector and the fidelity of expression of the pea vicilin gene in its new host was studied. The seeds of eight transgenic tobacco plants showed a sixteen-fold range in the level of accumulated pea vicilin. The level of accumulation of vicilin protein and mRNA correlated with the number of integrated copies of the vicilin gene. Pea vicilin was confined to the seeds of transgenic tobacco. Using immunogold labelling, vicilin was detected in protein bodies of eight out of ten embryos (axes plus cotyledon) and, at a much lower level, in two out of eleven endosperms. Pea vicilin was synthesized early in tobacco seed development; some molecules were cleaved as is the case in pea seeds, yielding a major parental component of Mr ~ 50,000 together with a range of smaller polypeptides.
Introduction

The seed storage protein fraction of most dicotyledonous plants is dominated by its globulin components. The globulins fall into two categories, the 11S legumin type and the 7S vicilin type. The most studied members of these two groups are the legumins from *Pisum sativum*, *Vicia faba* and *Glycine max* and the vicilins from the same species together with the vicilin-like protein (phaseolin) from *Phaseolus vulgaris*. Within each group, there is a significant level of sequence similarity between members from different plant species (7,29).

The seed storage globulins have many characteristics in common. They are all synthesized on the endoplasmic reticulum (ER) and are transported intracellularly via the lumen of the ER and the Golgi apparatus to be deposited in the vacuole or vacuole-derived protein bodies. An amino-terminal leader sequence is removed co-translationally during the translocation of the growing polypeptide chains into the lumen of the ER and all legumins and some vicilins undergo varying degrees of post-translational endoproteolytic cleavage after their deposition in the protein bodies (e.g. 11). The legumins are cleaved to yield two polypeptides which remain linked by disulfide bonds (22). The vicilins vary with regard to post-translational proteolytic processing. Phaseolin (from *P. vulgaris*) and ß-conglycinin (from *G. max*) are not substantially modified in this way (6,4), whereas vicilins from *P. sativum*, *V. faba* and *Phaseolus aureus* undergo extensive modification to yield a series of smaller polypeptides which form a part of the oligomeric proteins (see 22, for review). The biological function, if any, of these proteolytic cleavages of the original translation product is not understood.

This paper reports the isolation and sequencing of a vicilin gene from *P. sativum*. As part of a program to study the regulation of this gene and the fate of storage protein genes in a new host species, the vicilin gene was then stably introduced into the tobacco genome and a number of parameters of its expression were studied.
Methods

Construction and screening of a pea genomic library and sequencing of a vicilin gene

DNA was isolated from light-grown pea seedlings (10 days old) as described (37). The DNA was partially digested with Sau3A and fractionated by agarose gel electrophoresis and a size fraction corresponding to 14 to 20 kb was eluted from the low melting point gel (27). The DNA was ligated to BamH1-digested λ1059 arms packaged and propagated in E. coli strain C600 (25). A library of 10^6 plaque forming units was obtained and was screened with a pea vicilin cDNA clone pPS15-84 (36) as described (5). Three positive plaques were purified and characterised by restriction mapping. One (λ10.2) was subcloned into pUC8 (pEN2) and further characterised by sequencing using the dideoxy chain termination method (30). For this purpose the 5.5 kb EcoRI fragment was sonicated and the subfragments subcloned into M13 mp8 (14).

Transfer of the pea vicilin gene into tobacco

The plasmid, pEN2, was linearized at the unique BamH1 site in the polylinker and the entire clone was ligated into the BglII site of the Agrobacterium tumefaciens/E. coli shuttle vector pGA471 (1). This plasmid (pAGAB1) was transferred from E. coli (HB101) by conjugation to Agrobacterium tumefaciens (LBA 4404) using tri-parental mating methods (15).

Transconjugants containing the binary vector were selected on LB medium containing rifampicin (50 μg/ml) and tetracycline (5 μg/ml). The structure of the gene in Agrobacterium was verified by Southern blotting (34) and tobacco (Wisconsin 38) leaf pieces were transformed as described (23) with modifications. The leaf pieces were incubated with the Agrobacterium for 15 min and were transferred, without blotting and without feeder layers, to agar culture dishes containing basal MS medium (28). After two days the leaf pieces were transferred to agar culture dishes containing medium (MS9) for shoot induction (basal medium containing 0.5 mg/l indole acetic acid, 1 mg/l benzyl amino purine) and the antibiotics cefotaxime (300 mg/l) and kanamycin sulphate (100 mg/l). Shoots of about 1 cm in length formed within 21 to 28 days and were transferred to basal MS medium containing the same antibiotics for rooting. Plants that formed roots (10 to 14 days) were later transferred to soil in the glasshouse (26°C). Flowers were tagged at anthesis just as the petals opened and began
to turn a pink colour.

**Isolation and characterization of nucleic acids from tobacco**

DNA was isolated from young tobacco leaves as described by Sutton (1974). After digestion with restriction enzymes the DNA was fractionated on 0.8% agarose gels and transferred to nitrocellulose membrane (Schleicher and Schuell, Keene, N.H.). Hybridization was carried out using the 5.5 kb vicilin gene fragment from the clone pEN2. The fragment was labelled by nick-translation using a kit (Bresa, Adelaide, South Australia). Hybridizations were performed at 68°C without formamide in the buffer described by Singh and Jones (32). Heparin was replaced by polyanetholesulfonic acid, sodium salt (Sigma, St Louis, Mo.). Blots were washed for 2 x 10 min with 2 x SSC at 25°C, then for 3 x 20 min at 68°C in 2 x SSC containing 0.1% SDS and 0.1% sodium pyrophosphate and finally for 3 x 20 min in 0.1 x SSC 0.1% containing SDS and 0.1% sodium pyrophosphate at 60°C. The blots were fluorographed at -70°C using a Hi-Plus (Du Pont) intensifier screen.

Total RNA was isolated from frozen tobacco seeds by grinding in a mortar with a pestle using conditions described earlier (21) except that 1M Tris, pH 9.0 was used instead of 0.1M Tris. The integrity of the RNA was monitored by agarose gel electrophoresis in the presence of formaldehyde (27). S1 nuclease mapping was performed as previously described (38) using a single stranded DNA probe (24). The probe extended about 1000 bases upstream of position +210 (see Fig. 1). The protected fragments were separated on urea/acrylamide gels using, as size markers, 32P-labelled Mspl fragments of pBR322.

**Western Blotting**

Total protein was isolated by grinding fresh tissue with a mortar and pestle in 0.1 M N-Tris (hydroxymethyl)methyl-2-amino-ethanesulfonic acid (TES), pH 7.8, 0.2 M NaCl, 1 mM EDTA, 1 mM phenylmethylsulfonyl fluoride, 2% β-mercaptoethanol. The extracts were centrifuged at 10,000 x g for 15 min. The concentration of protein in the supernatant was measured using the Bradford reagent (8). Protein was fractionated by SDS-PAGE (35) and electrophorblotted onto nitrocellulose membrane as described (36). Vicilin was detected with primary antibodies
generated in rabbits using an immunodetection kit (Bio-Rad, Richmond, CA) which incorporates a second antibody linked to alkaline phosphatase for colorometric detection.

**Densitometry**

Fluorographs and negatives were scanned using a Joyce-Loebl double beam microdensitometer (MK111 CS). The relative contribution of individual bands was monitored by cutting and weighing the peaks.

**Immunocytochemistry**

Developing seeds from non-transformed (Wisconsin 38, W38) and transformed (B3, see Fig. 3) plants were prepared as described earlier (13). In brief, tissue was fixed at 0°C in 0.1% glutaraldehyde 2% paraformaldehyde for 1.5 h, post-fixed with 1% Os04, dehydrated through a series of ethanol solutions and embedded in LR White resin at 55°C.

Sections were treated with NaOH and HCl (12), prior to saturating non-specific binding sites of protein absorption with bovine serum albumin (10 mg/ml in 0.15 M NaCl, 10 mM Na-phosphate, pH 7.4 and 0.1% Tween-20). Anti-vicilin (20 µg/ml) and protein A-gold (15 nm) were used as described (13). All steps were carried out for 10 min at 22°C. Labelled sections were treated with uranyl and lead salts before viewing in a JEOL 100S electron microscope.

**Results**

**The structure of a pea vicilin gene**

A pea genomic library, constructed in λ1059, was screened with a vicilin cDNA clone, pPS 15-84 (kindly provided by Peter Chandler). Three positive clones were mapped by restriction endonuclease digestion and a 5.5 kb EcoRI fragment from one clone (λ10.2) was selected for detailed study. Part of the EcoRI subclone (4.2 kb of the 3'-end) was sequenced on both strands and the sequence was compared (Fig. 1) to both the pea vicilin cDNA clone (pPS 15-84) and to the gene sequence for an analogous 7S seed storage protein, phaseolin from *Phaseolus vulgaris* (16). Over 99% of the nucleotides in the coding region of the isolated pea
Fig. 1. The gene, cDNA and protein sequences of a pea vicilin. A phaseolin gene sequence (16) is included for comparison. \(a\) = protein sequence deduced from the vicilin gene; \(b\) = vicilin gene sequence; \(c\) = phaseolin gene sequence; \(d\) = vicilin cDNA sequence. \(\ast\) = gaps introduced to maximize nucleotide matching; \(\cdot\) = identity of bases. In line \(d\), a continuous line indicates the position of introns. Direct repeats in the vicilin sequence are overlined while the regions conserved between the vicilin and phaseolin genes around the "TATA" box and in the upstream region (−62 to −153) are boxed. The two downward arrows indicate the positions of post-translational processing in vicilin.

Note that because gaps have been introduced to maximize matching, the numbers refer to position only and not to actual nucleotide numbers in the sequence.
vicilin gene were identical to the vicilin cDNA clone used to isolate it (Fig. 1, lines b and d). The vicilin and phaseolin nucleotide sequences were aligned for maximal similarity by the placement of gaps in both sequences. As a result, the numbers used in Figure 1 refer to position rather than to consecutive nucleotides within either sequence.

The pea vicilin gene contained a direct repeat sequence in the 5' flanking region. It consisted of 54 bases (positions -466 to -519) which were repeated with only four mis-matches between positions -378 and -431 (note the two overlined regions in Fig. 1). No similar sequences were identified in the phaseolin gene at the 60% matching level.

The pea vicilin and phaseolin gene sequences showed an overall similarity of 39% in the sequences covering about 900 bp in the 5' regions plus the coding regions and about 150 bp 3' to the stop codons. However, there were short regions of much closer identity, particularly in certain parts of the 5' flanking sequences and in some of the exons. For example, there was notable sequence conservation in the 5' flanking sequences of the two genes between positions -200 and +1 (the cap site), particularly around the "TATA" box (position -33) and cap site and again further upstream, (the boxed positions between -153 and -62) covering a sequence of about 80 bases (Fig. 1). Although the 5' untranslated sequences were divergent there was sequence conservation in the 3' untranslated region and this was especially obvious upstream (starting at position 2620) of putative poly(A) addition signals which occur at positions 2570 and 2630 (Fig. 1).

Both vicilin and phaseolin genes contained five introns located at identical positions with respect to the amino acid sequence. The introns were all small but neither their precise size nor their sequence was conserved between the genes except around the splice junctions where they conformed to the consensus sequence described for other plant genes (9). Nucleotide sequence similarity in the six exons was highest in the third exon (70%) and lowest in the fourth (40%) and sixth (36%) exons of the gene. Sequence comparison at the amino acid level in general reflected the nucleotide comparison. Overall, there was 50% identity between the amino acid sequences of vicilin and phaseolin (data not shown), the third exon
showed highest conservation (63%) while the fourth and sixth exons were lowest (32 and 34%, respectively).

In pea seeds, some of the vicilin polypeptides undergo post-translational cleavage at one or both of two cleavage sites (19,36) while, in bean, phaseolin is not cleaved (33). The cleavage sites in vicilin are characterized by charged amino acids and cleavage occurs on the amino terminal side of an aspartic acid residue (note arrows at positions 1381 and 1996). When the vicilin and phaseolin amino acid sequences were aligned for maximum similarity the first processing site of vicilin was found to be deleted in phaseolin and the sequence around the second processing site was not conserved in phaseolin and there were six and twelve amino acid deletions in phaseolin on either side of the processing site (data not shown).

The expression of a pea vicilin gene in tobacco

A DNA construction containing the entire 5.5 kb EcoRI fragment of pEN2 together with a selectable neomycin phosphotransferase II gene was introduced into tobacco leaf pieces via an Agrobacterium tumefaciens binary vector transformation system. Kanamycin-resistant plants were regenerated from infected leaf pieces and the mature seed from these transgenic plants was analysed for the expression of pea vicilin by the Western immunoblot procedure (Fig. 2). Anti-serum against pea vicilin reacted with one major seed protein component (Mr ~47,000) in seeds from untransformed tobacco (Fig. 2, W38). In addition to this vicilin-related tobacco protein, anti-vicilin serum reacted with a number of other protein components in the seeds of the transformed tobacco (Fig. 2, B3). The major new component had an Mr of ~50,000 and co-migrated with the major component of vicilin from pea seeds (Fig. 2, vicilin, pea). A number of smaller cross-reacting polypeptides were also observed in the seeds of transgenic plants (Fig. 2, B3). The mobility of some of these polypeptides on SDS-PAGE is the same as that of the post-translational processing products of authentic pea vicilin (Mr ~50,000, 34,000, 27,000, 25,000 and 12,000) (Fig. 2, vicilin, pea) while others are unique to the transgenic tobacco seeds (Mr ~38,000, 19,000 and 10,000).

The level of pea vicilin expression in the mature seed of a range of nine independently transformed tobacco plants was examined by Western immunoblot analysis (Fig. 3). All plants
Fig. 2. Vicilin-related polypeptides in the seeds of non-transformed (W38) and transformed (B3) tobacco. Samples containing either 150 μg of protein from tobacco seed extracts, 3 μg of protein from pea seed extracts or 0.2 or 2.0 μg of vicilin purified from pea seed extracts were fractionated by SDS-PAGE, electroblotted onto nitrocellulose and probed with antiserum prepared against pea seed vicilin. Numbers on the vertical axes give the approximate $M_r \times 10^{-3}$ of a number of major vicilin-related polypeptides.
Fig. 3. The distribution of vicilin-related proteins in the seeds of nine putative transformants (B1 to B10) and the non-transformed progenitor (W38). Seed extracts containing approximately 350 µg of protein were fractionated by SDS-PAGE, electroblotted onto nitrocellulose and probed with antisera against pea vicilin. Numbers on the vertical axes indicate the approximate $M_r \times 10^{-3}$ of the polypeptides resulting from the expression of the introduced pea vicilin gene.
had been selected for kanamycin-resistance and this resistance was confirmed by at least one
further passage through tissue culture. Eight out of nine transformants contained detectable
levels of pea vicilin, with similar levels in plants B4, B6, B7 and B8 and progressively higher
levels in B10, B2, B1 and B3. One transformed plant (B9) failed to accumulate detectable
levels of pea vicilin.

An estimate of the amount of vicilin accumulated in seeds of the most highly expressing
transgenic tobacco (plant B3) was obtained by comparing densitometer tracings of a more
extensive series of vicilin concentrations than shown in Figure 2 with the level found in a
known amount of B3 seed protein. These indicated that pea vicilin comprised about 0.5% of
the total seed protein.

**Vicilin gene copy number**

DNA was isolated from the leaves of a selection of transformed tobacco plants, digested with
EcoRI endonuclease and analysed by the Southern blot procedure using the vicilin gene (5.5
kb EcoRI fragment of pEN2) as probe (Fig. 4). The level of pea vicilin DNA in the
transformants correlated closely with the level of accumulation of pea vicilin in the mature
seeds of the transformed tobacco plants. Transformant B9, which yielded no detectable pea
vicilin in the Western blot assay (Fig. 3), did not contain the pea vicilin gene but remained
kanamycin-resistant. Transformant B3, which gave the highest level of vicilin expression (Fig.
3), also contained the highest number of introduced genes. Densitometric comparison of the
signal obtained for the vicilin gene in the nine transformed tobacco plants with that given by a
set of gene reconstructions (not shown) indicated that, transformants B6, B7 and B10 each
have one to two copies of the introduced gene, B1 has 6-8 copies and B3 at least 30 copies. In
all cases studied there was a close correlation between the number of vicilin gene copies
detected and the level of vicilin protein in the transgenic tobacco. This suggests that in these
cases the actual position of insertion of the gene in the tobacco genome did not have a major
influence on the level of gene expression.
Fig. 4. The presence of the pea vicilin gene in six putative tobacco transformants (B1, B3, B6, B7, B9, B10). Total DNA (15 μg) from leaves of each tobacco plant and a sample of total pea DNA (15 μg) was digested with EcoR1, fractionated by electrophoresis on agarose, blotted onto nitrocellulose and probed with the vicilin gene (5.5 kb EcoR1 fragment from pEN2). Figures on the vertical axis show the size in kilobases of known DNA markers (λDNA digested with HindIII).
Intracellular localization of pea vicilin in tobacco seeds

Developing seeds of transformed (B3) and non-transformed (W38) tobacco were fixed, embedded and sectioned before challenging with anti-vicilin serum followed by gold-labelled protein A. Examination of these sections in the electron microscope revealed that vicilin-related proteins were accumulated in the protein bodies of the developing seed of transformed tobacco (Fig. 5a). In seeds of non-transformed tobacco there was a very low level of immunolabelling (Fig. 5B), presumably due to the vicilin-like tobacco proteins detected by Western blotting (Fig. 2, W38). The greatly increased level of labelling in both the cotyledon and embryonic axis of the seeds of the transformed plant reflects the accumulation of pea vicilin in these seeds. Tobacco protein bodies contain a distinct crystalloid structure within a background matrix. Both pea vicilin and the tobacco vicilin-related protein were largely confined to the matrix of the protein bodies (Fig. 5A,B).

Tobacco seeds differ from pea seeds in that they contain a significant volume of endosperm as well as cotyledon tissue in the mature seed. Although protein bodies in both storage tissues appear structurally similar (Fig. 5A,C), immunolabelling indicated that pea vicilin was largely confined to the cotyledons and embryonic axes of the transgenic tobacco seeds. Out of ten embryos (cotyledons plus embryonic axes) examined, vicilin was detected in eight cases. In contrast, it was detected, at a much lower level, in two out of eleven endosperm sections.

Organ specificity of vicilin gene expression

Western immunoblot analyses for vicilin in extracts of roots, stems, seeds and leaves of transformed tobacco indicated a high level of specificity in the expression of the vicilin gene (Fig. 6). In both transgenic tobacco and in peas, vicilin gene expression was confined to the seeds, although at much higher levels of loading of total protein, faint cross-reacting polypeptide bands were detected in stems and leaves of peas (Fig. 6, lanes 2 and 3) and in leaves of transformed tobacco (Fig. 6, lane 7).

Temporal pattern of vicilin expression in developing transgenic tobacco seeds

The patterns of pea vicilin accumulation in developing transgenic tobacco and of the
Fig. 5. Electron micrographs of protein bodies in cells of cotyledons (A) and endosperm (C) tissues from the tobacco transformant B3 and in cotyledon cells of non-transformed tobacco (B). Sections were immunogold-labelled for vicilin. A and C were recorded from a single section in which both cotyledon and endosperm tissue had been fixed *in situ*. Vicilin is sequestered within the proteinaceous matrix of the protein bodies but was not detected in the crystalloid, labelled (C). All micrographs are from developing seed at 19 days after flowering. Micrographs were printed to enhance the contrast of the colloidal gold. As a result the lipid bodies, which fill the adjacent cytoplasm, are not evident.
Fig. 6. The distribution of vicilin-related polypeptides in the organs of pea and transformed tobacco plants. Extracts of roots (lanes 1 and 5), stems (lanes 2 and 6), leaves (lanes 3 and 7) and seeds (lanes 4 and 8) were fractionated by SDS-PAGE, electroblotted onto nitrocellulose and probed with antiserum to pea vicilin. Lanes 2, 3, 5 and 6 were loaded with 150 μg protein, lanes 1 and 7 with 75 μg protein, lane 8 with 38 μg protein and lane 4 with 5 μg protein. Numbers on the vertical axis give the approximate Mr x 10^{-3} of vicilin-related polypeptides.
endogenous, vicilin-related protein of non-transformed tobacco were determined by Western immunoblot assay (Fig. 7). Vicilin-related polypeptides (Mr ~55,000) were detected in non-transformed seeds at the earliest harvest (11 days after flowering), although the particular polypeptide which is characteristic of mature seed (Mr ~47,000) was not detected until later in development (17 days after flowering). This transient appearance of higher molecular weight polypeptides earlier in development suggests a possible precursor relationship with the smaller, later product.

In the seeds of transgenic tobacco, pea vicilin (Mr ~50,000) was detected at the earliest harvest (11 days after flowering) and increased to a maximum proportion of total seed protein by 17 days after flowering, by which time all the cleavage products were evident. The early synthesis of the pea vicilin in tobacco suggests strongly that transcription of the vicilin gene is regulated in the same temporal fashion in transgenic tobacco as it is in peas (10).

This conclusion was supported by the changing levels of pea vicilin mRNA throughout development of transgenic tobacco seeds as measured by S1 protection assay (Fig. 8). Pea vicilin mRNA was detected at 11 days after flowering and the levels increased to a maximum at 15 (or 17) days and thereafter declined sharply to 19 and 21 days after flowering. Vicilin mRNA was not detected in non-transformed tobacco seeds (W38) at 17 days after flowering. The sizes of the two DNA fragments (183 and 180 bases) which were protected from the action of S1 nuclease by the mRNA were the same in both transgenic tobacco and in pea (Fig. 8) indicating that the start site(s) of transcription of the pea vicilin gene is/are the same in both these situations. From densitometric measurements, the maximum level of pea vicilin mRNA in the transgenic tobacco plant, B3, at 15 days after flowering was estimated to be approximately 1.5% of the level found in pea RNA from developing seeds at 15 days after flowering.

The fate of pea vicilin in transgenic tobacco seeds during germination

Pea vicilin in transgenic tobacco seeds was degraded early in seedling development (data not shown). Breakdown products of pea vicilin were apparent in transgenic tobacco seed extracts one day after imbibition and the original pea vicilin components were almost entirely degraded
Fig. 7. The pattern of accumulation of vicilin-related polypeptides during seed development in transformed (B3) and non-transformed (W38) tobacco. Seeds were harvested at intervals after flowering, extracts containing 150 µg protein were fractionated by SDS-PAGE, electroblotted onto nitrocellulose and probed with antiserum against pea seed vicilin. Numbers on the vertical axis show the approximate $M_r \times 10^{-3}$ of the major vicilin-related polypeptides.
Fig. 8. The level of vicilin mRNA in seeds of a transformed tobacco (B3) during seed development as determined by S1 protection assay. Lane 1 contains radioactive size markers generated by labelling pBR322 digested with Msp1 endonuclease. Lanes 4 to 10 contain 50 μg of total RNA isolated from developing tobacco seeds at 11, 13, 15, 17, 19, 21 and 17 days after flowering, respectively. The RNA was first hybridized to a 32P-labelled probe comprising 183 bases of the coding region and cap site together with a further one kb 5' to this region. The hybridized mixture was then digested with S1 nuclease and the products fractionated by electrophoresis on acrylamide gels and the protected fragments detected by autoradiography. Lanes 4 to 9 contain total RNA from seeds of a tobacco plant (B3) transformed with a pea vicilin gene, lane 10 contains RNA from a non-transformed plant (W38). Lanes 11 and 12 contain 1.0 and 0.1 μg total RNA, respectively, from developing pea seeds at 17 days after flowering. Lane 2 contains undigested 32P-labelled probe and lane 3 contains labelled probe and 50 μg tRNA digested with S1 ribonuclease.
after six days. This breakdown of pea vicilin followed the same time course as the breakdown of the vicilin-related tobacco protein (Mr ~47,000) during germination.

Discussion

We report here the isolation, DNA sequence and expression of a pea vicilin gene in a foreign host. Based upon restriction map comparisons, this gene probably belongs to the vicilin gene family at locus Vc-4 described by Ellis, et al. (17). This family of closely related vicilins (over 95% base matching) contains at least four members (P.M. Chandler, personal communication). Although it is not known whether this particular gene is expressed in pea seeds, it clearly has the potential to do so, because it can be expressed in the transgenic host, tobacco. The cDNA used to isolate this pea vicilin gene is over 99% identical, at the nucleotide level, to the corresponding sequence of the isolated genomic clone. Of the nine nucleotide changes in the coding region, six of these result in amino acid changes, some of which are not conservative in nature. No changes were found in the (short) 5' untranslated region and the cDNA contained a single 7 bp insertion in the 3' untranslated region. Both the isolated gene and the cDNA probe contained a putative polyadenylation signal (AATAAA) 31 bp upstream of the site of addition of the poly A tract, but there was a second putative signal further downstream in the gene (position 2634) which, if used, would result in two populations of mRNA with slightly different lengths of 3' untranslated regions. Lycett et al. (26) have reported the isolation and sequence of two partial cDNA clones (pDUB2 and pDUB7) encoding vicilin polypeptides. These clones showed 86% base similarity in their overlapping regions. The sequence of the vicilin cDNA reported here (pPS15-84) shows 84% and 85% base similarity to pDUB7 and pDUB2, respectively, where their coding regions overlapped. Thus, all three cDNA clones show substantial sequence divergence from one another, consistent with the findings of Ellis et al. (17) who showed at least three different loci for the vicilin genes in pea.

All genes for 7S seed storage proteins from a range of genera have similar basic structural features in that they all contain five introns (16). There are also strong nucleotide sequence similarities particularly within the coding regions of the 7S genes (7). Intron size and sequence are not conserved except for sequence conservation around the splice junctions. The 5'
flanking sequences of the 7S genes in pea, *Phaseolus vulgaris* and soybean are divergent except for three regions of conservation. Two are short sequences around the cap site and "TATA" box (16) and the third is a sequence of about 80 bp located about 100 bases 5' to the "TATA" box in all three genera but which occurred between positions -62 and -153 in the pea vicilin gene (16, 19, this study). The role, if any, of this "7S or vicilin box" in regulating gene expression, has yet to be determined.

When the pea vicilin gene was introduced into tobacco, it was expressed in its new host with a high degree of fidelity with respect to all the parameters studied. These included the temporal pattern of vicilin mRNA (Fig. 8) and vicilin protein (Fig. 7) accumulation during seed development, the start site of transcription (Fig. 8), the organ and tissue specificity of its expression (Fig. 5, 6), and its susceptibility to post-translational processing (Fig. 2). For several of these parameters, a similar high level of fidelity of gene expression was reported earlier when genes for two other vicilin-like proteins (phaseolin from *Phaseolus vulgaris* and β-conglycinin from *Glycine max*) were introduced into tobacco or petunia (2, 3, 31).

The preferential accumulation of the foreign protein in the cotyledons rather than the endosperm tissue of the transgenic tobacco seeds is particularly striking with both phaseolin (20, 31) and pea vicilin. Both these tissues of the tobacco seed contain protein bodies, but it appears that the pea vicilin and phaseolin genes contain regulatory sequences which specify embryo (cotyledon plus axis) expression. Only low and irregular accumulation of these foreign proteins was detected in the endosperm, which is maternal tissue. Using *in situ* hybridization, Barker *et al.* (2) have recently reported the strict localization of mRNA for soybean β-conglycinin in the embryo of transgenic tobacco.

The accumulation of newly synthesized vicilin in the protein bodies in transgenic tobacco seeds (Fig. 5) strongly suggests that the targeting signals inherent in the protein structure of vicilin elicit similar responses from the tobacco cell as from the pea cell. This in turn implies that, as is the case in pea (22), vicilin in transgenic tobacco is synthesized on the rough endoplasmic reticulum and transported via the Golgi apparatus to the protein bodies. However, we were unable to detect vicilin in either the endoplasmic reticulum or the Golgi
bodies of transgenic tobacco, presumably due to the low amounts present.

Some differences were observed in the detail of the post-translational cleavage of vicilin in transgenic tobacco in that not all the resultant polypeptides were the same size as in the pea seed (Fig. 2) and additional cleavage produced were detected. This implies either that the conformation of pea vicilin in tobacco differs from that in peas in a way that exposes other sites to proteolytic attack, or that proteases with different specificities are present in the tobacco seed protein bodies. A more extreme difference was seen in the case of phaseolin in transgenic tobacco (31). Phaseolin undergoes no major proteolytic processing in its natural host, but in transgenic tobacco about 70% of the primary translation product was cleaved to form a group of smaller polypeptides. However, the overall pattern of behaviour of pea vicilin and bean phaseolin in transgenic tobacco, such as accumulation in protein bodies and early breakdown during germination, all suggest that these “foreign” legume storage proteins fulfil a similar functional role in their new, non-legume host.

In peas, the vicilin polypeptides of Mr ~ 50,000 are the major products of a multigenic family (P.M. Chandler, personal communication). In earlier work on the post-translational processing of vicilin in peas (26,36), it was suggested that within the vicilin gene family some members had diverged sufficiently to generate different combinations of the two major processing sites, while other members had no processing sites. These latter would give rise to the major Mr ~ 50,000 polypeptides of mature vicilin. It was envisaged that if processing sites were present the molecule would be *totally* processed at that site. The behaviour of a single member of the pea vicilin family in transgenic tobacco contradicts this notion. The appearance of both a major Mr ~ 50,000 polypeptide and a series of smaller polypeptides as the product of a unique vicilin gene shows that partial processing of a single primary product does occur. In both pea and transgenic tobacco, the ratio between the primary translation product and the processing products is very consistent. This reproducible level of partial processing could be due to the fact that not all cleavable sites are accessible to proteolytic attack when vicilin is in its oligomeric (trimeric) configuration.
Several cases have now been recorded in which genes for 7S seed proteins from legumes have been transferred into tobacco and petunia (2,3,31 and the present study). In all cases these genes have been transcribed accurately from their correct start sites, introns have been spliced out of the gene transcripts and authentic protein products have accumulated. Furthermore, expression of these genes has closely paralleled that in their original host with regard to organ specificity and temporal regulation. This commonality of expression of these genes in their new hosts suggests that the flanking, regulatory regions of these genes could be used in chimeric gene constructions to ensure the effective organ-specific synthesis of "novel" seed storage proteins in grain legume crops.

Acknowledgements

We thank Dr. Gynheung An for supplying the shuttle vector (pGA 471), Dr. Peter Chandler for making available the unpublished sequence data on pea vicilin cDNA (pPS 15-84) and Dr. Christine Wandelt for comments during preparation of the manuscript. The technical assistance of Celia Miller and manuscript preparation by Denese McCann are acknowledged with thanks. E.J.N. is the recipient of a Commonwealth Post-graduate Scholarship.
References


