TRANSCRIPTIONAL REGULATION OF A PEA VICILIN GENE

A thesis submitted for the degree of

DOCTOR OF PHILOSOPHY

by

DIANA QUIGGIN

Department of Botany and Zoology
The Faculties
Australian National University
# TABLE OF CONTENTS

Declaration .................................................................................................................. v

Acknowledgments ......................................................................................................... vi

Abstract ........................................................................................................................ vii

Abbreviations ............................................................................................................... ix

Chapter 1. Literature review ........................................................................................ 1

1.1 Introduction .............................................................................................................. 1

1.2 Regulation of transcription in eukaryotes - general features ....................... 2
  1.2.1 Eukaryotic transcription is regulated by complex interactions between the
       DNA regulatory elements in a gene and the transcription factors that bind to
       these elements. ........................................................................................................ 2
       Transcription ........................................................................................................ 2
       Regulatory elements and transcription factors .................................................... 2
       Interactions between elements and transcription factors .................................. 4
  1.2.2 Modular structure of transcription factors. ...................................................... 7
       DNA-binding and dimerization domains .............................................................. 7
       Transactivation domains ...................................................................................... 8
  1.2.3 Regulation of transcription factors: What regulates the regulators? .......... 10
       Transcriptional and post-transcriptional regulation ............................................ 10
       Post-translational modifications ........................................................................ 11
       Protein-protein interactions ............................................................................. 12

1.3 An overview of regulation of gene expression in higher plants .................. 14
  1.3.1 Light-regulated gene expression ................................................................. 14
       The sequence motifs G and GT-1 are involved in light-regulated expression ..... 15
       Other light-responsive elements ..................................................................... 19

1.4 Transcriptional regulation of seed storage protein genes ......................... 20
  1.4.1 Seed storage proteins ..................................................................................... 20
       Protein classes .................................................................................................... 20
       Nomenclature ..................................................................................................... 21
       A common ancestral seed storage protein gene? ............................................. 22
  1.4.2 Temporal and tissue-specific control of globulin genes ......................... 22
       Seed storage protein genes are principally regulated by transcription ............... 22
       Temporal regulation ......................................................................................... 24
       Spatial regulation .............................................................................................. 24
       Seed-specific and temporal regulation are maintained within transgenic plants .. 24
       Are temporal and seed-specific regulatory elements really the same thing? ....... 26
       Seed-specificity may be regulated by positive or negative elements ................. 27
  1.4.3 Conserved DNA sequences in the 5' regions of seed-specific genes .......... 29
       Putative seed-specific DNA motifs .................................................................. 29
       Gene-specific elements ..................................................................................... 37
  1.4.4 Binding sites of putative transcription factors ............................................ 39
       DNA-protein binding assays ............................................................................. 39
       Many quantitative elements have AT-rich sequences that bind abundant
       nuclear proteins ................................................................................................. 42
       Functional analyses of AT-rich binding sites .................................................... 44
       Nuclear proteins bind to AT-rich sequences that may be structurally similar .... 45
       Nuclear proteins that bind AT-rich sequences fall into two classes ................... 47

1.5 Aim of thesis ............................................................................................................ 51
Chapter 2. Deletion analysis of the 5' region of Psa-v

2.1 Introduction

2.2 Materials and Methods

2.2.1 Construction of 5' deletion series of Psa-v

2.2.2 Transformation of plants

2.2.3 DNA isolation and Southern blotting of genomic DNA

2.2.4 Isolation and detection of vicilin in transgenic plants

2.3 Results

2.3.1 Analysis of transformed tobacco plants

2.3.2 Seed-specific expression

2.3.3 Temporal regulation

2.3.4 Delineation of quantitative cis-acting elements

2.4 Discussion

2.4.1 Seed-specific and temporal regulatory elements within Psa-v

2.4.2 Quantitative elements in Psa-v

Chapter 3. DNA-protein binding in the 5' region of Psa-v

3.1 Introduction

3.2 Materials and Methods

3.2.1 Preparation of crude nuclear extracts

3.2.2 Cloning of DNA fragments

3.2.3 DNA fragments used for probes and competitor DNA

3.2.4 Radioactive labelling of DNA fragments

3.2.5 Binding of the nuclear protein extract to the radioactive probe

3.2.6 Polyacrylamide gel electrophoresis

3.2.7 RNA isolation and northern blotting

3.3 Results

3.3.1 Temporal regulation of DNA-binding proteins

3.3.2 Location of binding sites in the seed-specific region and the minor quantitative element

3.3.3 Inhibition of DNA-binding proteins
3.4 Discussion
3.4.1 Organ specific and temporal regulation of DNA-protein binding
3.4.2 Comparison of DNA-binding sites in the seed-specific region of Psa-v with those in other 7S genes
3.4.3 AT-rich binding sites are prevalent in quantitative elements of seed-storage gene promoters
4.1 Introduction
4.2 Materials and Methods
4.2.1 Plasmids
4.2.2 Particle bombardment
4.2.3 Luciferase assays
4.2.4 GUS assays
4.3 Results
4.3.1 Assessment of specificity and temporal regulation of activity in transfected organs
4.3.2 Four regulatory DNA elements in the seed-specific region of Psa-v were identified by analysis of 5' deletion constructs
4.3.3 Transient expression of gene constructs with mutations within the region -306 to +39 of Psa-v
4.3.4 Functional Analysis of the TEMP-1 Binding Regions
4.4 Discussion
4.4.1 Transient expression of Psa-v-GUS constructs was temporally regulated and essentially seed-specific
4.4.2 The region -306 to +39 of Psa-v contains many regulatory elements
4.4.3 TEMP-1 binding regions confer high level expression in seeds
4.4.4 The role of CATGCATG motifs in seed-specific expression
4.4.5 Vicilin box I (octanucleotide and CACA) and box II sequences are located within positive regulatory elements
4.4.6 Negative and temporal regulation of Psa-v
4.4.7 Summary

Chapter 4. Functional analysis of the conserved DNA sequences and protein binding sites in the promoter of Psa-v
Chapter 5. Conclusions ............................................................................................................. 128

5.1 The 5' region of Psa-v contains many cis-acting elements and DNA-protein binding sites that are located in conserved or repeated DNA sequences. ........................................................................... 128

5.2 The binding activity of TEMP-1, TEMP-2 and TEMP-3 correlates with activation of Psa-v gene expression. ......................................................................................... 130

5.3 Activation of Psa-v expression in leaves may be regulated by a protein that inhibits DNA-protein binding of the "seed-specific" binding proteins ........................................................................... 131

5.4 How do the CATGCATG sequences affect gene regulation? ................. 132

5.5 Applications ..................................................................................................................... 133

Appendix I .................................................................................................................................. 135

Appendix II ............................................................................................................................... 140

Bibliography ............................................................................................................................ 155
DECLARATION

I declare that the results presented in this thesis are my own work and were 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 work of Dr Christine Wandelt who constructed the all the pCW plasmids used in chapter 2. Dr Edward Newbigin began construction of plasmids pQ1 and pQ3 which I used in the construction of pEN21 and pEN22, described in Appendix II. Dr Leon Dure assembled the majority of the 5' DNA sequences in the data base that was analysed in Table 1.2 and Appendix I.

Diana Quiggin

September 1993
Acknowledgments

Many people helped me at different stages of my study. To thank them all is not possible, but I will do my best.

I am extremely grateful to my supervisors TJ, Mark, Malcolm and Angela, especially TJ and Mark, who gave me advice, support and suggestions. Without them, this thesis simply would not have happened. I’d also like to thank Don for his advice and encouragement. Thank you Christine, Allison and Jim T. (and other people in Plant Industry) for patiently answering all my questions in the early days on plasmid making and gel-shift assays. Thank you also Audrey, for vital help slicing and cutting numerous pea seeds and tobacco capsules. John, Carol and Colleen were particularly noble for reading some of the more primitive drafts of parts of this thesis. In addition, I wish to thank the people who worked in the labs on level 3, both those present and those who have moved on, for creating such a lively and energetic atmosphere (even in the small hours), which made it such a great place to work.

My family also directly contributed to this work. I am particularly aware of my debt to my husband Bob, for so many things that it is impossible to list them, including being an excellent computing tutor and copy typist. My brothers Rohan and Jeremy provided invaluable help with the photography and graphics. Thanks also to my mother Meredith, to John and Nancy, and my parents-in-law Pat and Charlie for much babysitting of Patrick, and thank you Patrick for being so patient with your busy mother. Nancy and John also kindly let us use their house. Thanks to my father, Barry and to Audrey for proofreading and organising holidays at the coast. I am grateful to my whole extended family for their support and interest over a long period.

Johnno lent me a laser printer and a computer, (not to mention a CD player and lots of CDs) as well as assisting with the graphics. His generosity with his possessions, time and support was outstanding.

My friends inside and outside the lab have made the last few years very pleasant. Alison, Cathe, Geoff, Di, Mandy, Robbie, Victoria and Meredith kindly went swimming, cycling and walking with me and fed me on many occasions. Doug, Petra, Fiona, Kev, Vanessa and others all helped with last minute corrections and collation. Thank you to my other house mates, especially Peter and Sue, for putting up with me (and introducing me to the wide world of pickled and smoked everything). Others are not mentioned above, but all were indispensable. Thank you, everyone!

I would like to acknowledge the financial support given to me in the form of scholarship grants by the Commonwealth Government and the Australian National University, and my gratitude to Dr Peacock for making the facilities of the CSIRO Division of Plant Industry available to me.
Abstract

Seed storage proteins comprise the major proportion of protein in mature seeds. In pea, vicilin is usually the most abundant of these proteins, and is encoded by a small multi-gene family. Transcription of seed storage protein genes is virtually restricted to the seeds, and is confined to specific stages of development. The aims of this research have been to find regulatory DNA elements in the 5' region of the pea vicilin gene \textit{Psa-v}, to investigate the binding of nuclear proteins within these regulatory elements, and to analyse whether the conserved DNA sequences present in the 5' region of \textit{Psa-v} have a functional role.

Several DNA regulatory elements were located within the 5' region of \textit{Psa-v} (summarised in the diagram below). The expression of \textit{Psa-v} gene constructs, was analysed in both stably transformed tobacco plants and in transiently transformed pea organs. Regulatory elements which contributed to high-level expression in seeds were located, using 5' deletion and internal mutation analysis, between -773/-425, -424/-307, -150/-118, -117/-90, -73/-62 and -61/-4. Other positive elements were located between -185/-151 and -89/-74. Many of these positive elements contained conserved DNA motifs that are common to genes that are expressed in seeds, indicating that these motifs may have a role in the transcriptional regulation of the \textit{Psa-v} gene.

The \textit{Psa-v} gene constructs were expressed at low levels in transiently transformed leaves and petals. Deletion of -117/-90 increased expression in these organs. Thus this region, which contains a CATGCATG motif, may contribute to seed-specific expression by both enhancing expression in seed tissue and repressing expression in non-seed tissue. Another negative element, -230/-194, was also identified, but deletion of this sequence increased expression in seeds as well as in petals and leaves, indicating that it may act as a more general negative element.

Six of the regulatory regions that were defined were shown to contain sequences that were specifically recognised by nuclear proteins extracted from developing pea seeds (see diagram). Regions containing binding sites for four seed-specific nuclear proteins, TEMP-1, -2, -3, and AT-2 were determined by gel-shift and competition assays. As well as correlating with major regulatory elements, these binding sites also contained conserved DNA motifs or repeated DNA sequences. AT-2 bound to sequences within -512/-486 and
-424/-398, which both contain two repeats of the sequence GATAa/gATAAT. Three regions required for binding of TEMP-1 were located, within -185/-151, -152/-118 and -115 /-96. These protein-binding sequences contain vicilin box II, the 5' region of vicilin box I (containing an octanucleotide motif) and the 3' region of vicilin box I (containing a CACA motif), respectively and have the consensus sequence TTtttTtcTtTTTgacacat. TEMP-2 and TEMP-3 both bound within -61/-4, which contains the TATA box. TEMP-1, -2 and -3 binding activities were temporally regulated, peaking at developmental stages that corresponded with expression of Psa-v, suggesting that these proteins may have a major role in the temporal and seed-specific expression of this gene.

None of these four seed nuclear proteins were detected in leaf nuclear protein extracts. By mixing seed, leaf and other nuclear extracts, it was shown that the leaf nuclear extract contained a factor that inhibited both the binding of nuclear proteins to the Psa-v promoter and a protein that recognises AT-rich sequences in LegA1, another pea seed storage protein gene. This factor had little or no effect upon the binding of either of two other DNA-binding proteins; Octapine Synthase Element-binding protein or Epstein Barr Virus Nuclear Antigen-1. The leaf extract also contained other binding proteins that interacted with Psa-v DNA fragments. This indicates that the inhibition by the factor was specific to certain DNA-binding proteins, suggesting that transcription of Psa-v may be repressed in leaf tissue, to some extent, by the inhibition of binding of trans-acting factors to binding sites within the 5' region.

### 5' Region of Psa-v

<table>
<thead>
<tr>
<th>Conserved DNA Motifs</th>
<th>Regulatory Elements</th>
<th>Binding of Nuclear Proteins</th>
</tr>
</thead>
<tbody>
<tr>
<td>AT-rich tandem repeats</td>
<td>box I</td>
<td>TEMP-2</td>
</tr>
<tr>
<td>-773</td>
<td>box I</td>
<td>TEMP-3</td>
</tr>
<tr>
<td>-424</td>
<td></td>
<td>AT-2</td>
</tr>
<tr>
<td>-306</td>
<td></td>
<td>AT-2</td>
</tr>
<tr>
<td>-231</td>
<td></td>
<td>-</td>
</tr>
<tr>
<td>-194</td>
<td></td>
<td>+</td>
</tr>
<tr>
<td>-150</td>
<td></td>
<td>+++</td>
</tr>
<tr>
<td>-117</td>
<td></td>
<td>-</td>
</tr>
<tr>
<td>-89</td>
<td></td>
<td>+</td>
</tr>
<tr>
<td>-61</td>
<td></td>
<td>+</td>
</tr>
<tr>
<td>+1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---
ABBREVIATIONS

Abbreviations used often in this thesis are:

A  adenine nucleotide
C  cytosine nucleotide
G  guanine nucleotide
T  thymine nucleotide
ABA  abscisic acid
ATF/CREB activating transcription factor / cAMP response element binding protein
bp  base pairs
bZIP  protein with a basic DNA binding domain and a leucine zipper dimerization domain
cab  chlorophyll a/b binding protein gene
cAMP  cyclic AMP
CaMV  cauliflower mosaic virus
CAT  chloramphenicol acetyl transferase
cDNA  copy DNA
EDTA  ethylenediaminetetra-acetic acid
GUS  β-glucuronidase
HMG  high mobility group
kb  kilobase pairs
MAB  MAB Gene Gun (gunpowder driven gun)
mRNA  messenger RNA
NOS  nopaline synthase
PIG  Particle Inflow Gun (helium driven gun)
rbcS  ribulose-1,5-bisphosphate carboxylase-oxygenase gene
SDS  sodium dodecyl sulphate
SSC  standard saline citrate
Tris  tris(hydroxymethyl)aminomethane
CHAPTER 1. LITERATURE REVIEW

1.1 Introduction

The development and growth of plants and animals relies upon the precise spatial and temporal control of gene expression. Complex organisms begin as single cells and grow by the replication, differentiation and expansion of those cells. The coordinated regulation of the genes within these cells, and the subsequent actions and regulation of the proteins that these encode, are what ensure the orderly development of the complex variety of specialised organs, tissues and cells that comprise a higher plant or animal.

The seed has evolved in higher land plants to maximise the survival potential of the succeeding sporophyte generation. After fertilisation, the seed begins to develop, forming new structures, principally the embryo and the endosperm. The embryo differentiates to form one or two cotyledons, and shoot and root meristems.

In most legume and some other dicotyledonous species, the initial, liquid endosperm is absorbed during embryo development in the seed. In cereals, however, the endosperm becomes solid, and accumulates seed storage proteins that will act, as in legume cotyledons, as a source of energy and nutrients during and after germination. The seed matures, dehydrates and becomes biochemically quiescent until suitable germination conditions ensue.

All these developmental steps are mediated by the activation and repression of different sets of specific genes and their products. The genes that encode seed storage proteins are expressed specifically in the seed. In addition, the expression of these genes is generally restricted to the earlier stages of seed development (reviewed in Higgins, 1984; Goldberg et al., 1989). This seed-specific and temporally-regulated expression of seed storage protein genes makes them an appropriate system for the study of regulatory mechanisms of gene expression in higher plants. The research in this thesis was aimed at analysing the regulatory mechanisms of one such gene.
This chapter reviews the general features of the regulation of eukaryotic transcription and then discusses the regulation of particular plant genes, especially those encoding seed storage proteins.

1.2 Regulation of transcription in eukaryotes - General features

1.2.1 Eukaryotic transcription is regulated by complex interactions between the DNA regulatory elements in a gene and the transcription factors that bind to these elements.

TRANSCRIPTION

A eukaryotic gene consists of three functional parts; the coding region and the 5' and 3' non-coding regions (Figure 1.1). The 5' region usually contains the "cis-acting" DNA elements which, in interaction with "trans-acting" nuclear protein factors regulate the transcription of the gene (Figure 1.2). The part of the 5' region proximal to the start site of transcription, called the promoter, contains sufficient information for the initiation of transcription of the gene. Regulatory elements upstream of the promoter region, often termed upstream activating sequences or enhancer elements, are involved in enhancing the rate of transcription, limiting it to certain cell types and responding to specific cellular signals. When a gene is activated or induced, it is transcribed into messenger RNA by RNA polymerase II. This primary transcript is then processed by the addition of a methylated nucleotide cap at the 5' end and, usually, a poly-A tail to the 3' end. The introns, or intervening sequences, are also excised. The mature transcript is then transported from the nucleus to the cytoplasm and translated into protein (Figure 1.3).

REGULATORY ELEMENTS AND TRANSCRIPTION FACTORS

The 5' regions of eukaryotic genes generally contain an assortment of short DNA sequences (6-12 bp) called cis-acting elements. Many of these regulatory elements have been shown to contain binding sites for specific proteins called transcription factors (Figure 1.1). Regulatory elements fall into two basic classes; common and gene-specific (reviewed (Maniatis et al. 1987; Dynan, 1989)).
Figure 1.1
Diagram of a gene and its transcriptional control region.

A. A gene consists of three regions: 5', coding and 3' sequences. Transcription starts upstream of the coding sequence and is controlled by DNA elements in the 5' region.

B. The 5' region contains a proximally located promoter and distally located quantitative elements. The promoter region is composed of a TATA sequence and other short DNA-sequences elements (boxes) which are the binding sites for specific transcription factors. The TATA box is located 20-30 bp upstream of the transcriptional start site and is the DNA-recognition site for the transcription complex, which includes the general transcription factors (TFIID and others) and RNA polymerase II (pol II). Both the proximal promoter and distal enhancer elements contain the DNA-binding sites for numerous, different gene-specific transcription factors (shown as boxes and ovals respectively). As indicated by the arrows, these factors act in a combinatorial manner to modulate the activity of the transcriptional complex. (Figure adapted from Peterson and Baichwal, 1993).
Common regulatory elements

The most important element common to genes transcribed by RNA polymerase II is the TATA box (reviewed in Peterson and Balchuf (1993)). It is found 20 to 30 bp upstream from the transcription start site and is required for efficient and accurate transcription by RNA polymerase II. The transcription factor TFIID binds specifically to the TATA box. It is thought that RNA polymerase II is then anchored at the initiation site by the sequential interaction of TFIID with other factors TFIIA, TFIIIB, RNA polymerase II and finally TFIIIE (Bjørn et al., 1989).

A.

Other common elements act as binding sites for transcription factors, such as the CAAT and GC boxes, which are critical for the initiation and expression of genes (reviewed in Maniatis et al., 1987).

Gene-specific elements

Transcription of most tissue-specific or inducible genes is mediated by the binding of specific DNA-binding proteins to their recognition sites, and the subsequent interaction of the bound proteins with the initiation complex which then activates or represses transcription by RNA polymerase (Figures 1.1 and 1.2). Examples of gene-specific elements are those that regulate expression in response to extracellular stimuli, environmental factors, and developmental signals.

B.

Some of these elements have been shown to be bound by DNA-binding proteins that activate gene expression. The cyclic-AMP response element (CRE) confers inducibility by cyclic AMP and is activated in a number of transcription factors in the transcription factor / cAMP response element binding protein (activating transcription factor / cAMP response element binding protein) families (reviewed in Maniatis et al., 1987). Heat shock response elements, which are functionally conserved in Drosophila and man, can complex with heat shock transcription factors. Purified heat shock element-binding protein has been shown to activate transcription in vitro (reviewed in Maniatis et al., 1987).
Common regulatory elements

The most important element common to genes transcribed by RNA polymerase II is the TATA box (reviewed in Peterson and Baichwal (1993). It is found 20 to 30 bp upstream from the transcription start site and is required for efficient and accurate transcription by RNA polymerase II. The transcription factor TFIIID binds specifically to the TATA box. It is thought that RNA polymerase II is then anchored at the initiation site by the sequential interaction of TFIIID with other factors TFIIA, TFII B, RNA polymerase II and finally TFIIE (Buratowski et al., 1989).

Other common elements act as binding sites for transcription factors, such as the CAAT and GC boxes which are recognised by the CTF (CAAT binding transcription factor) and Sp1 ("promoter-specific" factor) binding proteins, respectively (reviewed in Mitchell and Tjian (1989), Kadonaga et al. (1986).

Gene-specific elements

Transcription of most tissue-specific or inducible genes is moderated by the binding of specific DNA-binding proteins to their recognition sites, and the subsequent interaction of the bound proteins with the initiation complex which then activates or represses transcription by RNA polymerase (Figures 1.1 and 1.2). Examples of gene-specific elements are those that regulate expression in response to extracellular stimuli, environmental conditions and developmental signals.

Some of these elements have been shown to be bound by specific DNA-binding proteins that activate gene expression. The cyclic-AMP response element, for example, confers inducibility by cAMP and is also implicated in calcium- and virus-induced transcriptional regulation. This element is recognised by the many binding proteins in the ATF/CREB (activating transcription factor / cAMP response element binding protein) families (reviewed in Lamb and McKnight (1991). Similarly, heat shock response elements, which are functionally conserved between yeast, Drosophila and man, can complex with heat shock transcription factors. Purified heat shock element-binding protein has been shown to activate transcription in vitro (reviewed in Maniatis et al. (1987).
Transcriptional regulation through interactions between DNA-binding proteins

In 1.2.1, transcription is activated by the binding of factor A and the interaction of this factor with the transcription complex. This transcription is further enhanced by the presence of extra elements, and the subsequent interactions between the additional bound proteins (factor A and/or enhancer proteins, B) and the transcription complex.

Factors A and B can be abundant and able to activate transcription in cell-types a and b, respectively, but unable (independently) to bind and activate transcription in cell-type c, due to low concentration of each factor in c cells. The presence of both recognition elements may result in gene expression in cell-type c, due to co-operative binding between A and B as dimers.

Transcription can be repressed by the binding of a repressor protein (B) to a defined site, perhaps at some distance from target promoter, which interferes with the action of the activator.

Activation by transcription factor A can be indirectly competed by the binding of a repressor protein at or near the transcription site, which blocks the interaction of general transcription factors such as RNA polymerase. This activation can also be directly competed, whereby the repressor binds to, or directly adjacent to the binding site of factor A, preventing A from binding.

Activation can also be "quenched" by the binding and interaction of proteins at two adjacent DNA sequences, where protein-protein interactions between A and B prevent A from making proper contact with the transcription complex. This could be caused by either masking or induction of a conformational change in the transactivation domain of factor A, when bound by factor B.

Another form of transcriptional repression, squelching, can occur by interaction of the activator in solution with another protein (not necessarily another DNA-binding protein) which results in the sequestration of the activator protein.
1 ACTIVATION

2 REPRESSION

3 COMPETITION

4 DIRECT COMPETITION

5 QUENCHING

6 SQUELCHING
INTERACTIONS BETWEEN ELEMENTS AND TRANSCRIPTION FACTORS

Many different types of interaction occur between regulatory elements and their binding. Such interactions can enhance or repress expression in a given cell type or result in expression in a new cell type (Figure 1.2).

Enhancement

Regulatory elements are often conserved between genes that are expressed in the same tissues or respond to the same environmental or developmental stimuli. In many cases gene-specific elements are also repeated within individual promoters and these extra copies may have an additive effect, enhancing the level of gene expression (Figure 1.2.2). For example, metallothionein genes from humans all contain multiple copies of a 12 bp regulatory element, and the level of heavy metal induction for each gene is proportional to the number of copies present in the 5' upstream region (Stuart et al., 1984; Stuart et al., 1985). Additive effects have also been noted for other elements including heat shock and glucocorticoid regulatory elements reviewed in Maniatis et al. (1987). Other regulatory elements, termed enhancers, upstream activating sites or (positive) quantitative elements can also increase the level of transcription of a gene through the interaction of activating proteins (see Guarente, 1988; Dynan, 1989 for review). Transcriptional activation by some proteins, such as the mammalian DNA-binding protein Oct-1 (Schaffner, 1989), may also require the presence of an adaptor protein to act as a bridge between the DNA-binding protein and the target DNA site (see Ptashne and Gann, 1990, for review).

Synergistic interactions

Although the combination of regulatory elements often has an additive or enhancing effect, it does not always follow that the combination of different regulatory elements has a simple modular or additive effect upon transcriptional regulation. Combinations of different elements within a promoter can have a synergistic or an antagonistic effect, resulting in the activation or repression of gene expression (see Miner and Yamamoto, 1991, for review).

One example of synergistic expression between regulatory elements, leading to expression in new tissues, has been demonstrated for the promoter of the cauliflower mosaic virus (CaMV) gene encoding 35S RNA. The intact promoter confers high levels of expression in most tissues. This promoter is composed of seven individual regulatory regions, each conferring different expression...
patterns in plants, but the sum of these expression patterns of each domain, in the different tissues and during development, does not equal that conferred by the entire promoter. Furthermore, combinations of regulatory regions resulted in expression patterns that were different from those of the individual regions. For example, fusion of sub-domain B1 to domain A conferred strong expression in the cotyledons of both seeds and seedlings even though these domains alone conferred little or no expression to this tissue (Benfey and Chua, 1990).

It has been proposed (Frankel and Kim, 1991), that gene expression in a new cell type, by the combination of two regulatory elements that do not usually activate expression in this cell type, could be mediated by co-operative binding of the respective transcription factors to the regulatory elements, resulting in activation of expression (see Figure 1.2.3).

Repression and Competition

Expression in a novel tissue, conferred by the presence of regulatory elements that do not usually confer expression in this cell-type, can also be due to the displacement of a negative element within the promoter. An example of repression of gene expression, by the interaction of two elements (Figure 1.2.4) has been shown for a vascular-specific cell wall protein gene from Phaseolus vulgaris. This tissue-specificity is maintained by the presence of a negative element that represses the activation of expression of a positive element in non-vascular tissue. This repression may be mediated by the binding of a repression protein, which then prevents the action of a transcription factor at the positive element. Furthermore, this repression was disrupted by insertion of 33 bp between the negative and positive elements (Keller and Baumgartner, 1991) which illustrates that the distance between regulatory elements can also affect the interaction between elements.

Repression of transcription can also be caused by the binding of a repressor protein at, or adjacent to, the transcription start site, so that although the activating factor can still bind to its recognition site, it is unable to transactivate expression (Figure 1.2.5). Probably the most common form of transcriptional repression, however, is where a transcription factor is in direct competition for a binding site (Figure 1.2.6). For example; the binding of the interferon regulatory element IRF-1 is competed by the binding of IRF-2, to the same site which subsequently represses expression of type 1 interferon genes (Kuhl et al., 1987). Similarly, binding of a CCAAT-binding factor is prevented by the binding of a tissue-specific displacement protein to the CAAT box in a histone
gene (see Levine and Manley, 1989; Falvey and Schibler, 1991; Renkawitz et al., 1991; for review).

**Quenching**

Antagonistic interaction between elements, resulting in a loss of expression has also been noted, and may be as a result of "quenching" (Figure 1.2.7). Quenching of expression by the interaction of two factors, which may otherwise activate transcription, could be caused by steric hindrance, so that when both factors were bound to their recognition sites, the proximity of the factors interferes with, or masks the regions involved with the activation of transcription (trans-activation domains). Alternatively, interaction between the two binding proteins could cause conformational changes to the transactivation domains, which would also prevent activation.

The polyoma virus enhancer, for example, contains a salivary/kidney-specific element whilst the Moloney murine leukaemia virus contains a lymphoid-specific enhancer. A chimeric gene containing both these elements, however, was not expressed in any of these tissues. The enhancer was active within the pancreas instead, indicating that the interaction of these two elements was both antagonistic and synergistic (Rochford et al., 1987). The yeast mating type protein α2, probably prevents the transactivation of the MCM1 activator protein on a-specific genes by quenching. The repressor binds at an adjacent site to MCM1 and so it may act by steric hindrance, masking the activation domain of MCM1, or alternatively by inducing a conformational change in the MCM1 protein (see Levine and Manley, 1989 and references therein).

**Squelching**

Sequestration of the binding protein, by interaction with other proteins in the cytoplasm or nucleus, can also prevent binding of a protein to its recognition site (Figure 1.2.8). This term, inelegantly termed "squelching", occurs when a transcription factor, A, is bound by another protein, B, which inhibits binding of A to its target site.

Thus, there are many mechanisms by which the combination of cis-acting elements, and the interaction of the transcription factors that bind to them, could determine the specificity and level of gene expression.
1.2.2 Modular structure of transcription factors.

In general, transcription factors consist of two essential domains; a DNA-binding domain and a transactivation domain. Some factors also contain a domain that is used to form homo- and hetero-dimers. The DNA-binding and transactivation domains are physically separate and function independently, indicating that transcription factors have a modular structure with different region for different functions (extensively reviewed in Mitchell and Tjian, 1989; Struhl, 1989; Frankel and Kim, 1991; Latchman, 1990; Peterson and Baichwal, 1993).

DNA-BINDING AND DIMERIZATION DOMAINS

Transcription factors are classified into families based upon the DNA-binding and/or multimerization domains. Different types of DNA-binding domains have been identified. Whilst some proteins bind as monomers, many factors bind as dimers and in these cases the DNA binding domain is often located close to a dimerization domain.

Two classes of factor that bind as monomers do so via either a homeodomain or a zinc finger structural motif. The homeodomain was first identified as a protein segment of about 60 amino acids that is conserved within proteins that regulate segmentation during embryogenesis of Drosophila. Homeodomains are also present in mammalian transcription factors. As well as being highly conserved, these domains are not actually restricted to factors that regulate developmental expression; for example one is present within the octamer binding protein, Oct-1, which participates in the expression of many ubiquitously expressed genes, including histone genes (Levine and Hoey, 1988).

Zinc finger motifs are present within the DNA-binding site of Sp1 and many other regulatory proteins, including a serum-inducible factor, Krox-2 (Nardelli et al., 1991; Mitchell and Tjian, 1989). The DNA-binding domain of steroid hormone receptors contains two zinc fingers. These differ structurally from the former zinc finger class. Steroid hormone receptors bind initially as monomers, then dimerize (Mitchell and Tjian, 1989; Beato, 1991; Peterson and Baichwal, 1993).
Other transcription factors, that bind as multimers, belong to the bZIP (basic DNA binding domain plus leucine zipper dimerization domain), helix-loop-helix and rel-related families. Many transcription factors contain a bZIP binding domain, including the yeast transcriptional activators GCN4, and proteins of the ATF/CREB and related oncogenic Fos/Jun families. These bind as either homodimers (for example Jun/Jun) or heterodimers (for example Fos/Jun). Dimerization is selective, however; for example Fos cannot dimerize with other Fos proteins. Mutational studies of a Jun/Fos pair demonstrated that the zipper motif is required for the formation of dimers, whilst mutations in the basic region block DNA-binding, but not dimerization (Struhl, 1989; LaLamb and McKnight, 1991).

A second major class of transcription factors that dimerize rely upon a common domain that probably forms a helix-loop-helix structure. Proteins in this class include Myc and its partner Max (Blackwood and Eisenman, 1991) and the muscle differentiation-determining factor, MyoD1 (Davis et al., 1987; Benezra et al., 1990). It was originally thought that these proteins bound to the DNA via the helix-loop-helix structure, but, like the bZIP proteins, deletion studies of MyoD1 found that the dimerization domain was an adjacent basic motif (Lamb and McKnight, 1991; Latchman, 1990; Peterson and Baichwal, 1993).

Not all binding domains have been elucidated, including a family that has a conserved but unknown DNA-binding domain, the rel-related proteins. These proteins, which include the lymphocyte activation factor NF-κB and the oncogenic rel, have a role in the induction of gene expression in differentiated cells (Lamb and McKnight, 1991; Peterson and Baichwal, 1993).

**TRANSACTIVATION DOMAINS**

The role of a transactivation domain is to communicate with the general transcription machinery, either to repress or to activate transcription in response to environmental, developmental or other stimuli. These domains may interact with the TATA-binding protein complex, or another protein within the transcription initiation complex, or they may alter the structure of the surrounding chromatin to permit other proteins to bind nearby.

Functional analysis has defined different transcriptional activation domains. The major transactivation domains within Sp1 are two glutamine-rich regions. CTF, on the other hand, contains a proline-rich domain which activates transcription when linked to various DNA-binding domains (Mitchell and Tjian,
Transcriptional activation domains in yeast transcription factors such as GCN4 and GAL4 are contained within short, negatively-charged, acidic regions that can form amphipathic α-helical structures (Mitchell and Tjian, 1989). The transactivation domain of Myb, an oncogenic transcription factor, is also acidic (Luscher et al., 1990). The acidic domains of GAL4 and ATF have been shown to alter the conformation of TFIID when it is bound to the TATA box, thereby facilitating the binding of other factors such as TFIIB and TFIIE as well as RNA polymerase II, to form a stable initiation complex (Latchman, 1990). These acidic transactivation domains may interact with one or more of the proteins of the initiation complex.

Since a range of structural motifs can activate gene expression, it will be interesting to learn whether these different motifs activate transcription in the same way (that is, via the transcription complex), or whether the type of transactivation domain is also a potential means for the fine tuning of transcriptional regulation.

Transcription factors may contain multiple, different transactivation domains. For example, GAL4 contains two acidic activation domains, whilst Sp1 contains two other domains (a basic domain and a COOH-terminal domain) in addition to the two glutamine-rich domains (Mitchell and Tjian, 1989). The presence of multiple and varied transactivation domains obviously increases the number and type of potential interactions that a transcription factor could engage in, thus permitting more complex regulation.

Thus, not only can DNA binding and dimerization be accomplished by a variety of structural domains, but transactivation can also be achieved by different types of structural domains. The physical structure of each DNA-binding domain probably constrains the DNA sequences that can be recognised. Similarly, the individual transactivation domains may be limited by the way in which they can activate gene expression. Therefore, combinations of both DNA-binding and transactivation domains (and dimerization) probably diversify the DNA sequences that are recognised and so contribute to the intricate regulation of the transcription of eukaryotic genes.
Figure 1.3
A transcription factor can be regulated at many stages.

The diagram depicts possible steps that might regulate the abundance or activity of a transcription factor. They begin with production in the nucleus of the mRNA of a transcription factor, followed by the transport of the mRNA into the cytoplasm, translation into protein, and transport of the protein into the nucleus to its site of action. Any individual factor might be modulated at only one of a few of these points. For examples, see the text. (Figure was derived from Figure 1 from Falvey and Schibler, 1991).
::::{·:\i•:·,:(-/: :.•:·'.,·.

::=.·.'·:··•·········.'.;;·:.i.·.f,,.·.·

·=::<{>?::::

.

,)<;'····

: :;~~:.

..

Transcription
factor gene

t

-::><X XXXX )C.

transcription

5

P~iRNA

,

, ;/
,,

3'

I

IY ~r~~esslng
mRNA

'

adaptor protein

~ ~:•

..,._"""""-.,...:_..

t
~~\

!i~~i~:.~\

.~

l
!

Complex
formation
with additional
protein(s) in
nucleus

5' - - - - - - - 3'

degradation

........

""f,.,,
J,~

...,.,

. ,~,,,~

••.f··

''.1,,,,,,
' {",,,1f#.q,,

.., ..,,,

:•1:;::"•••.....,.......'.¥''''"'·''''''~'"''"'"'1"''"'':••"''' .

••

:<l(::,

,

,. ,

•, '-:· ·::: ~;;
•'·\\•

:,.

-.;.

.;~ ~~;l~~i oi~ a~ofr;,
:v,,ith ac{ditlon~
Prot'ein_{s)
n,cyt9plasm
\,

,.._:::.:

:..,·.

\.' '' _\:J~::,

,:.

':.-:

,1

~~~+J~


1.2.3 Regulation of transcription factors:
What regulates the regulators?

When hunting for the DNA elements and protein factors that regulate a gene, one should consider how the regulatory factors themselves might be regulated. Like any enzyme, the activity of a transcription factor can be regulated either by controlling the level of abundance of the factor itself, or by altering the activity of the factor. Figure 1.3 shows various mechanisms by which the abundance or activity of a transcription factor could be regulated. These fall into four basic categories: (1) regulation of transcription, (2) regulation of translation, (3) protein modification and (4) protein-protein interactions. Investigation into the regulation of transcription factors has shown that each of these types of mechanism is employed, for different gene systems (see Falvey and Schibler, 1991; Hunt, 1989; Levine and Manley, 1989; Latchman, 1990; Foulkes and Sassone-Corsi, 1992 for reviews).

TRANSCRIPTIONAL AND POST-TRANSCRIPTIONAL REGULATION

A direct method of regulating the abundance of a factor and ensuring its presence in the right cells at the right time, is through the regulation of transcription of the gene that encodes it. Many cell-type and tissue-specific transcription factors are controlled at this level (see Latchman, 1990; Falvey and Schibler, 1991). One example includes the factor Oct-2, which is involved in the regulation of 8-cell specific transcription of immunoglobulin genes (Muller et al., 1988; Scheidereit et al., 1993; Clerc et al., 1988).

Cell-specific or tissue-specific transcriptional regulation of a factor, however, implies that there must be other regulators controlling the expression the gene encoding this factor, suggesting a cascade of transcription factors. Such regulation, therefore, does not solve the problem of gene regulation, but simply sets it at one further remove.

Post-transcriptional mechanisms have been elucidated, however, that modulate the synthesis of specific transcription factors in response to primary signals. For example, translation of GCN4, is generally blocked by a ribosome stalled at the end of a small open reading frame. This yeast factor activates the expression of genes encoding for enzymes required for amino acid biosynthesis. Induction by amino acid starvation relieves the ribosomal block and results in the increased translation of the GCN4 messenger RNA (Roussou et al., 1988). Such a mechanism, regulating at a post-transcriptional level,
would permit faster induction than one involving transcriptional control, and so might be useful for the rapid induction of a gene in response to environmental stimuli.

There are many other steps between the production of a primary transcript and the mature polypeptide that could moderate the abundance or activity of a transcription factor. These include intron removal and RNA splicing, poly A tail addition, transport to the cytosol, modulation of the translation rate and altered mRNA stability. Of these steps, transcription factors have been shown to be regulated by alternative RNA splicing (Foulkes et al., 1991; Nakabeppu and Nathans, 1991; Foulkes et al., 1992; also see Foulkes and Sassone-Corsi, 1992 for review) and initiation of translation at alternative start sites (Descombes and Schibler, 1991).

POST-TRANSLATIONAL MODIFICATIONS

Once a transcription factor has been synthesised, its activity can be regulated by modifications to the DNA binding, dimerization or transactivation domains.

Phosphorylation of transcription factors has been shown to both inhibit and activate different DNA binding and transactivational domains within transcription factors (exhaustively reviewed by Hunter and Karin, 1992). For example, the DNA-binding activities of the transcription factors cMyb, Max, Oct-1 and c-Jun are repressed by phosphorylation. On the other hand, phosphorylation has been shown to stimulate the transactivation domains in the factors CREB, cMyc, GAL4, c-Jun, and a yeast heat shock factor. (reviewed by Falvey and Schibler, 1991; Hunter and Karin, 1992).

Phosphorylation can also have a role in determining the location of a transcription factor within a cell. In addition to transcriptional and post-transcriptional regulation, the relative abundance of a transcription factor (in the nucleus) can be regulated by the exclusion of the factor from the nucleus. One transcription factor regulated in this way is NF-κB. This transcription factor is involved with the activation of κ chain gene transcription in B lymphocytes. NF-κB resides in the cytoplasm of unstimulated cells as a complex with an inhibitor, I-κB. Induction of cells with phorbol esters induces the activity of protein kinase C, which phosphorylates the inhibitor, rendering it unable to bind to NF-κB, which is then able to move to the nucleus (Baeuerle and Baltimore, 1988a; Baeuerle and Baltimore, 1988b; Ghosh and Baltimore,
In this example, the protein that is modified is the inhibitor, rather than the transcription factor itself. The regulation of NF-κB is also an example of how phosphorylation can affect dimerization.

Protein modifications other than phosphorylation can activate the DNA-binding site of a transcription factor. The presence of copper causes a conformational change in ACE1, a yeast factor which activates transcription of the metallothionein gene, which permits ACE1 to bind to the regulatory sites within the gene and to activate transcription (Furst et al., 1988).

**PROTEIN-PROTEIN INTERACTIONS**

The activity of transcription factors can also be regulated by interactions with other proteins, before and after binding to the regulatory DNA element. These protein-protein interactions include the formation of hetero- and homodimers with similar transcription factors, as well as interaction with dissimilar proteins, and can affect either DNA binding or transcriptional activation domains.

Dimerization can generate an extension in the already diverse repertoire of transcriptional regulation by altering the specificity and/or affinity of a transcription factor for its target. Members of the bZIP family bind to their DNA sites as dimers; Jun members can form homodimers, whilst Fos members cannot form homodimers, but form heterodimers with Jun transcription factors. Both Fos and Jun dimers bind to DNA sequences known as AP-1 sites, which mediate induction following phorbol ester treatment. Hetero-dimerization can create new affinities and binding specificities. Jun-Jun homodimers bind to AP-1 sites, but with relatively low affinity. In contrast, Fos-Jun heterodimers bind to the same site with much higher affinity, reflecting a difference in the stability of the dimers. Whilst these Fos-Jun heterodimers bind to the AP-1 site in preference to a related cAMP response element, the reverse is the case for Jun-ATF heterodimers (reviewed in Jones, 1990; Latchman, 1990; Lamb and McKnight, 1991).

Heterodimerization can also repress the activity of a transcription factor; the binding of different proteins, to form a heterodimer instead of a homodimer, can inhibit the transactivation function (Blackwood and Eisenman, 1991; Descombes and Schibler, 1991; Foulkes et al., 1991; Nakabeppu and Nathans, 1991; Foulkes et al., 1992).
Binding by a protein can also anchor a transcription factor within the cytoplasm (as shown with NF-κB). In the case of MyoD1, the DNA-binding function is repressed by heterodimer formation with an inhibitory protein, Id. This transcription factor is required for the initiation of muscle-specific genes during cell differentiation, but activation of gene transcription with the differentiating cells occurs via a decrease in the level of Id, rather than an increase in the concentration of MyoD1. Id and MyoD are structurally similar, and can dimerize via helix-loop-helix domains. Since Id lacks an adjacent basic DNA-binding region, the activation of gene expression is probably repressed by the formation of heterodimers that are incapable of binding to the regulatory site (Davis et al., 1987; Benezra et al., 1990).

Other forms of protein-protein interaction that also affect the activity of a transcription factor, have been described above (p4), and are depicted in Figure 1.2. Such interactions: repression, quenching, squelching, enhancement and synergism, can regulate binding to a recognition site, or the activation of the transcription complex by a bound factor.

The regulation of the activity or abundance of a transcription factor need not be moderated at just one step; for example, the gene encoding Pit-1/GHF-1, which is a pituitary-specific factor that controls the expression of growth hormones (Bodner et al., 1988; Ingraham et al., 1988) appears to be regulated by many factors, including an unknown specific pituitary transcription factor, cyclic AMP levels and positive autoregulation by the Pit-1/GHF-1 protein itself (McCormick et al., 1990; Chen et al., 1990; Dolle et al., 1990).

Also, one regulatory step can have more than one effect upon the regulation of a transcription factor. For example, the production of both full length and truncated transcription factor by differential RNA splicing or initiation of translation at alternative sites, firstly regulates the abundance of the transcription factor by reducing the number of full length factors that are present. The second, and apparently, stronger effect also reduces the level of active transcription factor, but via heterodimerization between the full-length and truncated proteins (reviewed in Foulkes and Sassone-Corsi, 1992).

An abundance of mechanisms exist, therefore, that can regulate the activity of a transcription factor. These include control of accumulation and alteration of affinity and activity of the mature factor. Accumulation can be regulated at the transcriptional, post-transcriptional and intracellular levels. The activity and
affinity of a transcription factor can also be regulated at the DNA-binding, dimerization and transactivation domains by a vast variety of different protein-protein interactions or by modification of the protein. It may be that the types of mechanisms that are used to regulate a given transcription factor reflect the speed and intensity of response that is requires when the activity / abundance of a transcription factor is induced or repressed.

1.3 An overview of regulation of gene expression in higher plants

Although most of the current knowledge of transcriptional regulation of eukaryotic genes is derived from work on yeast, *Drosophila* and mammalian cells, considerable progress has been made in plant gene research. The ability to transfer genes into host plants using the *Agrobacterium* Ti system (Horsch *et al*., 1984) has been extremely useful for the analysis of gene regulation. Transient expression systems (utilising electroporation of cells, or microparticle bombardment of tissue) have also proved useful. Many different types of plant genes have been analysed, especially those that are developmentally-regulated or are tissue-specific (including seed-specific genes, reviewed by Edwards and Coruzzi, 1990; Bewley and Marcus, 1990; Goldberg *et al*., 1989; Chee and Slightam, 1991; de Lumen, 1990), or are induced by various environmental conditions (reviewed by Gilmartin *et al*., 1990a; Kuhlemeier *et al*., 1987).

1.3.1 Light-regulated gene expression

Transcriptional regulation in plants has probably been most extensively studied in light-regulated genes. Since the knowledge of light-regulated genes is more detailed than that for seed-expressed genes, it is outlined below, to give examples of the types of transcriptional regulation that have been observed in plant genes.

Functional *cis*-acting elements within the promoters that regulate expression of light-regulated genes, and the DNA-binding proteins that specifically recognise these *cis*-acting elements, have been identified. Furthermore, some of the genes encoding these nuclear proteins have been isolated and characterised. The major families of light-responsive genes that have been investigated are those encoding small subunit of Rubisco, ribulose-1,5-bisphosphate
carboxylase-oxygenase (rbcS) and chlorophyll a/b binding protein (cab).
Broglie et al. (1984), using a pea rbcS gene in petunia, were the first of many to demonstrate that the cis-acting, light-responsive elements from one species could be recognised by the transcription factors from another species. Since then substantial advances have been made in locating cis-acting sequences within light-regulated genes, determining the trans-acting factors that bind to them, and how they are regulated.

**THE SEQUENCE MOTIFS G AND GT-1 ARE INVOLVED IN LIGHT-REGULATED EXPRESSION**

A number of reiterated and redundant DNA sequences have been found in the upstream regions of rbcS and cab genes. These conserved sequences include the G-box and GT-1 which contain the binding sites of the putative transcription factors GBF (G-box factor) and GT-1, respectively.

G elements are present in cis-acting elements of many plant genes.

The G-box, CACGTG, was originally detected as a conserved sequence in rbcS and cab genes, but also appears to be conserved and active in the 5' region of many other types of plant genes. G-boxes are associated with high transcriptional activity in rbcS and cab genes (Castresana et al., 1988; Donald and Cashmore, 1990; Luan and Bogorad, 1992) and may have a regulatory role in response to UV light (Schulze-Lefert et al., 1989a; Schulze-Lefert et al., 1989b).

A protein, GBF has been identified that binds in vitro to the G box within tomato, pea and Arabidopsis rbcS genes (Giuliano et al., 1988) and the Nicotinia plumbaginifolia cab-E gene (Schindler and Cashmore, 1990). GBF is present in nuclear extracts from both light-grown and dark-adapted tomato plants, but there are differences in the abundance and the mobility of the complexes formed by this DNA-binding protein from the different extracts. GBF binding activity is greater in extracts from light-grown plants than in dark-adapted plants (Fiebig and Link, 1992; Giuliano et al., 1988). In addition, the DNA-protein complexes that form between the G-box (Giuliano et al., 1988) or G-box-like motifs (Fiebig and Link, 1992) and nuclear proteins from dark-adapted plants have faster mobility than those formed with nuclear proteins from light-grown plants, suggesting that the binding proteins may be modified in response to light.
Three genes encoding G-box factors, called GBF1, GBF2 and GBF3, encoding bZIP proteins, have been isolated from Arabidopsis (Schindler et al., 1992a). DNA binding activity of GBF1 is stimulated 10- to 20-fold following phosphorylation by broccoli casein kinase II. Phosphorylation of GBF 1 also reduces the mobility of the resultant DNA-protein complex in polyacrylamide gels (Klimczak et al., 1992), similar to that formed with extracts from light-grown plants. Thus, the activity of GBF may be regulated, in response to light conditions, by controlling both the abundance of GBF and by protein modification.

The G-box is also conserved within other light-regulated genes. Two cis-acting elements, box II and box III (not related to GT-1 boxes, below), have been located within the light-responsive chalcone synthase promoter in parsley. Box II, ACGTGGG, within one of these elements is very similar to the G-box. Three cDNAs have been isolated encoding proteins that bind specifically to box II. At least one of the common plant regulatory factor proteins, CPRF-1, appears to be regulated at the transcriptional level rather than at the protein level; mRNA levels increased after exposure of dark-adapted plants to light. These CPRF proteins, like GBF1, also belong to the bZIP class of transcription factors and bind to the core motif ACGT within the G-box (Weisshaar et al., 1991).

The G-box, however, does not appear to be limited to light-responsive gene promoters. G-box motifs are present within the 5' regions of many different types of genes including those that encode seed storage protein genes, vegetative storage proteins, alcohol dehydrogenases and heat shock proteins (Williams et al., 1992). Also, many cis-acting elements have been identified that contain the core binding motif of the G-box, ACGT. These elements include the Em1a motif, which is involved in the response to abscisic acid by the wheat Em gene (Marcotte et al., 1989), the hex1 motif from the wheat histone H3 promoter which is thought to play a role in cell cycle-dependent transcription (Tabata et al., 1991) and the as-1 element from the CaMV 35S promoter (Lam et al., 1989). Genes encoding bZIP proteins that bind to these motifs have also been isolated; EmBP-1 (Guiltinan et al., 1990); HBP-1 (Tabata et al., 1991).

The bZIP proteins CPRF-1, CPRF-2 and CPRF-3, bind to the ACGT-containing promoter elements in Em1a, hex1 and as-1, as both homodimers and heterodimers, though with markedly different affinities (Armstrong et al., 1992). Although all plant bZIP proteins that have been examined to date bind to the core ACGT (Katagiri and Chua, 1992), they can be separated into classes that
bind in the presence of different flanking sequences (Williams et al., 1992; Schindler et al., 1992b). Since the bZIP proteins and their binding sites appear to be involved in the regulation of many genes, it may be that specificity is mediated, like the Fos/Jun bZIP transcription factors (page 8), through the formation, binding specificity and interactions between particular dimers. Specific expression patterns of bZIP protein genes in different organs may also be important in regulating tissue-specific and inducible regulation.

**GT-1 elements**

The GT-1 elements are present in most *rbcS* promoters and have been separated into two more conserved sequence types; box II and box III, which share a common consensus sequence and which, although not identical, act as GT-1 binding sites (see Thompson, 1991). These binding sites are usually present as multiple (and to some extent redundant) copies within the promoters of *rbcS* and *cab* genes. The 5' region of the pea gene *rbcS-3A*, for example, contains six GT-1 binding sites, including boxes II and III, and further upstream, related sequences within boxes II*, III*, II** and III** (Green et al., 1987; Green et al., 1988). The *cab-E* gene in *Nicotiana plumbaginifolia* also contains multiple GT-1 binding sites. These seven GT-1 binding sites do not formally fit into box II or box III classifications but all share a loose consensus with the *rbcS-3A* sites over a 6 bp sequence (Schindler and Cashmore, 1990).

Analyses of boxes II and III showed that mutations of these boxes resulted in the loss of GT-1 binding *in vitro* and also the loss of transcription *in vivo* (Green et al., 1988; Kuhlemeier et al., 1988).

Genes encoding three proteins that bind to GT-1 sites have been cloned and characterised from different plant species (Dehesh et al., 1990; Gilmartin et al., 1992; Perisic and Lam, 1992). The proteins they encode do not contain DNA-binding domains that make previously identified motifs, but appear to share structural similarities with each other. Predicted amino acid structures of the rice protein, GT-2 (Dehesh et al., 1990) and tobacco protein, GT-1a, both contain a novel helix-helix-turn-helix structural motif that appears to contain the DNA binding domain (Gilmartin et al., 1992). Tobacco B2F (Box II factor) also contains three putative α-helices that may also be involved in DNA binding (Perisic and Lam, 1992). GT-1a and B2F contain no known transactivation domains, whereas GT-2 contains an acidic domain (Dehesh et al., 1990; Gilmartin et al., 1992; Perisic and Lam, 1992).
Further studies of the functional role of the box II sequence indicated that the presence of the box could indeed activate transcription, presumably via binding by GT-1, but that interaction with an additional protein was required (Davis et al., 1990; Lam and Chua, 1990). GT-1 is present and active in extracts from both light-grown and dark-adapted plants (Green et al., 1988) which also suggests that an additional protein, one that is regulated in response to light, may be required for transcriptional activation. Alternatively, the activity of GT-1 could be regulated in vivo, by binding of a dark-specific protein, which could prevent binding and/or transcriptional activity (see Figure 1.2.4 and 1.2.5), or, as for NF-κB (p1), limit the location of GT-1 to the cytoplasm, in dark-adapted plants.

In support of the hypothesis that other binding proteins, (see figure 1.3) in addition to GT-1, are required for the transactivation of rbcS-3A, recent analysis by Sarokin (1992) has shown that two other proteins bind within the promoter. Mutational analysis revealed that box III contains two short sequences that are critical for expression and that these sequences serve as binding sites for proteins termed 3AF5 and 3AF3. Gel shift assays using DNA fragments with each of these sites showed that complexes with faster mobilities were formed using nuclear extracts prepared from dark-adapted plants compared with those from light-grown tobacco plants. It has been proposed that these differences in mobility were due to the different phosphorylation status of the proteins in the two extracts, since phosphatase treatment of the extracts from light grown plants resulted in the formation of complexes with faster mobility. One limitation of these experiments, however, was that it was not reported whether kinase treatments could retard the mobility of the proteins in the dark-adapted extracts. Binding of 3AF3, but not 3AF5, was dependent upon the (putative) phosphorylation status. This indicates that some aspects of light-regulated expression might be mediated by 3AF3; perhaps the phosphorylated forms of 3AF3 and GBF, which are present in the light-grown plants, binds to their recognition sites, then interact with prebound GT-1 and 3AF5, and the resulting complex then activates transcription.
OTHER LIGHT-RESPONSIVE ELEMENTS

In addition to the GT-binding sites and the G-box, other DNA binding motifs are preserved within light-responsive genes. These include the 3AF1 binding site and the multiple GATA (or I box) and AT-1 binding sites (reviewed in Gilmartin et al., 1990) and ABF-2 binding site (Arguello et al., 1992). The DNA binding activity of pea AT-1 can be reversibly modulated by phosphorylation. Unlike GBF-1 and 3AF3, it is active in the non-phosphorylated form and loses all DNA binding ability as a result of phosphorylation (Datta and Cashmore, 1989).

Light-regulated genes, therefore, contain an abundance of cis-acting elements, including the GT-1, G-box, 3AF1, 3AF3, 3AF5, GATA and AT-1 binding sites. The proteins that bind to these sites belong to various families of nuclear proteins, including a novel set of GT-1 binding proteins containing a DNA-binding domain with a helix-helix-turn-helix amino acid motif, the bZIP G-box binding proteins, and a zinc finger 3AF1 binding protein (Lam et al., 1990a). There is evidence that interaction between at least some of these proteins is required for effective transcription. Thus although many proteins bind to the promoter region of light-responsive genes, transcriptional activation may be dependent upon binding of certain crucial light-regulated proteins factors. This could explain why the activity of some of the DNA-binding proteins, such as GT-1 and GBF, are not light-specific; they might moderate activity but not activate it.

Thus, the activity and/or abundance of the nuclear proteins that bind to light-responsive promoter elements appear to be regulated in different ways. Some may interact with other binding proteins (eg. GT-1); others are regulated at either transcriptional/translational or post-translational levels in response to light conditions. The abundance of CPRF-1 and GBF-1 appears to be reduced in dark-adapted tissue whilst the DNA binding ability of 3AF, AT-1, and perhaps GBF1 appears to be regulated by light-dependent post-transcriptional phosphorylation. Differences between the mechanisms of transcriptional and post-translational regulation would affect the rate at which a given DNA-binding protein could respond to induction by light, and may have a bearing upon the role of the different proteins.
1.4 Transcriptional regulation of seed storage protein genes

**INTRODUCTION**

In the past few years, numerous genes encoding seed storage proteins have been isolated. Three approaches have been used to find DNA sequences in these genes are responsible for the seed-specific, temporal and quantitative regulation of seed storage protein gene

- sequence comparison; to find conserved DNA motifs
- functional analysis of gene constructs in transformed plants and cells; to determine the location of important regulatory elements, and
- DNA-protein binding assays; to find sites within the 5' regions of these genes that are recognised by nuclear DNA-binding proteins.

The spatial and temporal regulation of each of the four classes of seed storage proteins described below is different. Also, some of the conserved 5' regions that have been identified are specific to certain gene classes. So, although the seed storage protein genes, and other seed-expressed genes may be regulated by some common conserved DNA elements and transcription factors, it is likely that other regulation may be specific to certain genes or gene classes. Consequently, the following sections focus largely, but not exclusively, upon the transcriptional regulation of genes from legumes encoding vicilin and legumin proteins.

**1.4.1 Seed storage proteins**

**PROTEIN CLASSES**

Seed storage proteins comprise the major proportion of protein in mature seeds. Seed proteins can be divided into four classes of globulins, albumins, prolamine and glutelins, which are soluble in salt, water, alcohol, and acid or alkali solutions (Osborne, 1924). Prolamines are the predominant proteins in most cereal seeds. In some cereals such as oats (Avena sativa) and rice (Oryza sativa), however, the major storage proteins are globulins (Brinegar and Peterson, 1982) and glutenins (Yamgata et al., 1982), respectively.
The major storage proteins in legumes and other dicotyledonous plants are globulins and, to a lesser extent, albumins. Globulins are divided into two families (which have different mobilities in a sucrose gradient): the glycosylated 7S vicilins and the non-glycosylated 11S (and 12S) legumins. In addition to the 2S albumin seed storage proteins, the albumin fraction of seeds contains many other proteins. Some of these represent constitutive house-keeping enzymes that are essential for normal cell maintenance and function. Other albumins include lectins and proteinase inhibitors. Although these are expressed highly in the seed of some species, they are generally not seed-specific. In addition to acting as defence proteins against pests and diseases, lectins and proteinase inhibitors function as seed storage proteins (reviewed in Higgins, 1984; Goldberg, 1986; Edwards and Coruzzi, 1990). Other proteins that are highly expressed in seeds include those associated with seed maturation/desiccation and germination such as late embryogenesis abundant (lea) proteins in cotton (Baker et al., 1988; Dure et al., 1989a) and the aleurone-specific α-amylases in wheat (Lazarus et al., 1985), barley (Huang et al., 1984) and rice (Huang et al., 1990).

NOMENCLATURE

The names of seed storage proteins, and the constituent polypeptides of which they consist, can be confusing. The trivial names for the 7S (vicilin) proteins from pea (Pisum sativum) are vicilin and convicilin; from soybean (Glycine max), β- and γ-conglycinin; from French bean (Phaseolus vulgaris), phaseolin and from jack bean (Canavalia ensiformis), canavalin. Similarly, the 11S (legumin) proteins from pea, soybean and sunflower (Helianthus annuus) are termed legumin, glycinin and helianthinin, respectively. Legumin and vicilin consist of multimers with 6 and 3 polypeptides, respectively, and in some cases Greek prefixes have been used to discriminate between these polypeptides, for example; α- and β-legumin; α-, α'-, and β-,β'-conglycinin and α-,β-,γ- and δ-phaseolin (Casey et al., 1986). The trivial names for the prolamines include zein of maize (Zea mays), hordein of barley (Hordeum vulgare), secalin of rye (Secale cereale) and gliadin of wheat (Triticum aestivum). The glutelins of wheat are termed glutenins.
A COMMON ANCESTRAL SEED STORAGE PROTEIN GENE?

There are many structural similarities between the different seed storage proteins, which suggests that the gene families that encode them may have arisen from a common ancestral gene. Structural analysis of globulins from legumes shows that the vicilins and legumins share a common protein domain, which is duplicated within vicilin proteins (Gibbs et al., 1989). Legumins also strongly resemble the rice glutelins (Takaiwa et al., 1987a). Other glutelins are structurally similar to the prolamin proteins and are now considered to be members of the prolaminate class (Shewry and Tatham, 1990). In addition, there are structural similarities between prolamin and 2S albumins (see Okita et al., 1989; Shewry and Tatham, 1990). Since the genes coding for these proteins may have evolved from a common precursor, this suggests that these genes may share common, conserved regulatory elements.

1.4.2 Temporal and tissue-specific control of globulin genes

SEED STORAGE PROTEIN GENES ARE PRINCIPALLY REGULATED BY TRANSCRIPTION

The seed-specific and temporal expression of storage protein genes is primarily under transcriptional control, although post-transcriptional processes modulate the final amount of the gene products. Messenger RNA transcripts for seed storage protein genes accumulate to high levels during early- and mid-maturation stages of embryogenesis and then decrease to low or undetectable levels by the late-maturation stage (Chandler et al., 1984; Walling et al., 1986, Harada et al., 1989; Nielsen et al., 1989). The cells of non-embryonic organs, such as leaves, stems and roots, contain little or no detectable seed storage protein gene transcripts (Goldberg et al., 1981; Gatehouse et al., 1982; Walling et al., 1986).

The changes in the steady state levels of these messenger RNAs during embryogenesis and negligible levels within vegetative parts of the plant could be because the genes are transcribed at different rates and/or the stability of the mRNA changes, during different temporal stages and in different tissues. In vitro transcription assays, using isolated nuclei, show that transcription of globulin genes is activated and then repressed during embryogenesis, and furthermore that these genes are not transcribed, or are very weakly transcribed, in other organs of the plant (Beach et al., 1985; Harada et al., 1989; Nielsen et al., 1989; Walling et al., 1986; Evans et al., 1984). These results
indicate that gene transcription is the major point at which spatial and temporal regulation of seed storage genes occurs.

Post-transcriptional events, however, can influence the timing and final amount of the accumulated seed storage proteins, although they are not known to influence their seed-specific expression. Differences between the steady-state levels of mRNA species of globulin genes with similar rates of transcription indicate that post-transcriptional events also regulate globulin mRNA levels (Walling et al., 1986; Harada et al., 1989; Nielsen et al., 1989). In many cases, however, this may be a result of passive post-transcriptional regulation. This sort of regulation is observed when, of two transcripts present in the same cell, one is more susceptible to degradation by ribonucleases, perhaps because of a less stable secondary structure.

Some seed storage protein genes have been shown to be actively regulated at the post-transcriptional level; that is, they contain regulatory sequences in the 3' region that result in the degradation of the mRNA under special circumstances. For example, the genes encoding legumin and pea albumin 1 in pea (see Rerie et al., 1992), are regulated by nutritional conditions. Growth under sub-optimal sulphur conditions reduces the accumulation of legumin and albumin in the seed. This is due to increased turnover of the mRNAs, rather than a change in the rate of transcription of these genes (Chandler et al., 1983; Higgins et al., 1986).

Some 7S proteins may be weakly transcribed in non-seed tissues (Beach et al., 1985; Walling et al., 1986; Higgins et al., 1988). Hence reduced stability of the mRNA in vegetative organs might also contribute to the very low concentration of seed storage proteins in these tissues. Furthermore, seed storage proteins contain amino acid "signals" that direct them to the vacuoles, which in the leaves, are rich in proteases (Chrispeels, 1991). Thus, seed storage proteins are more rapidly degraded in leaves than in seeds (Lawton et al., 1987; Higgins and Spencer, 1991; Bagga et al., 1992; Wandelt et al., 1992).
TEMPORAL REGULATION

Seed storage proteins do not accumulate at the same time during seed maturation. In legumes, the 7S family of globulins are generally detected before those of the 11S family. In pea, for example, the level of vicilin gene transcripts increases rapidly between 10 and 13 days after flowering, these levels peak at about 16 days after flowering before declining to undetectable levels by 25 days after flowering. In contrast, legumin gene transcripts are not detected until 2 to 3 days after the appearance of the vicilin mRNAs and their level also peaks and declines at much later stages of seed maturation (Chandler et al., 1984). The patterns of accumulation of the seed storage proteins directly correlate with the relative abundance of their corresponding mRNAs during seed development (Meinke et al., 1981; Gatehouse et al., 1982; Bassuner et al., 1983; Walling et al., 1986; Yang et al., 1990).

SPATIAL REGULATION

In addition to organ specificity there is tissue specificity within the seed itself. In particular, the seed storage proteins of legumes are predominantly restricted to the embryonic tissues (Goldberg et al., 1981; Meinke et al., 1981; Walling et al., 1986; Ladin et al., 1987; Pang et al., 1988; Perez-Grau and Goldberg, 1989), whilst those of cereals are restricted to the endosperm (reviewed in Shotwell and Larkins, 1989).

This spatial-specificity, however, is not absolute. For example, legumin (Harris et al., 1989) and probably vicilin (Czako et al., 1992) genes are expressed in the transient, liquid endosperm. Globulins are also capable of expression in non-embryonic tissue in cereals, as shown by the vicilin-like proteins of barley endosperm (Yupsanis et al., 1990) and wheat aleurone (Quatrano et al., 1988; Singh et al., 1988). It seems unlikely, therefore, that the differences in tissue specificity are due to crucial structural differences between prolamines and globulins, but rather that the genes that encode them are regulated differently.

SEED-SPECIFIC AND TEMPORAL REGULATION ARE MAINTAINED WITHIN TRANSGENIC PLANTS

Gene transfer studies have shown that the DNA sequences that control seed-specific and temporal expression are contained within the 5' regions of seed storage genes. Many intact seed storage protein genes have been shown to be correctly regulated in transgenic plants; expression of these genes is both seed-specific and temporally regulated (see Goldberg et al., 1989; and Table
Mature tobacco seed contains both endosperm and embryonic tissue and so can be used to study the expression of both legume and cereal seed storage protein genes. Seed-specific and tissue-specific regulation is generally preserved by promoters from either embryo-specific or endosperm-specific genes in transgenic tobacco. Many legume storage protein genes (or promoters of these genes ligated to reporter genes), when tested in transgenic tobacco, were detected solely in embryonic tissue (Sengupta-Gopalan et al., 1985; Barker et al., 1988; Bustos et al., 1989). In some cases, a low level of expression has also been detected in the endosperm (Greenwood and Chrispeels, 1985; Higgins et al., 1988; Sturm et al., 1988; Bogue et al., 1990; Baumlein et al., 1991a; Baumlein et al., 1991b; Czako et al., 1992). These results support the findings that legumins (Harris et al., 1989) and vicilins (Czako et al., 1992) are expressed in the endosperm during very early stages of seed development.

Tissue-specificity of most cereal storage genes is also maintained in transgenic tobacco (Colot et al., 1987; Marris et al., 1988; Schernthaner et al., 1988; Robert et al., 1989; Matzke et al., 1990; Quattrochio et al., 1990; Takaiwa et al., 1991b). Interestingly, some of the cereal genes do not maintain seed-specificity; a zein gene (Ueng et al., 1988) and a reporter gene with a rice glutenin promoter (Leisy et al., 1989), were both expressed in leaves as well as seeds. This suggests that not all the regulatory elements within cereal genes are recognised appropriately by the tobacco transcription factors.

Despite the probability that seed storage protein genes may have evolved from a common precursor tissue, and so may share some common consensus elements, because dicotyledonous and monocotyledonous seed storage protein genes are expressed primarily in different tissues of the seed, it seems unlikely that all the elements controlling their expression are identical. Also, some of the transcription factors that recognise these cis-acting elements may have diverged among species.
ARE TEMPORAL AND SEED-SPECIFIC REGULATORY ELEMENTS REALLY THE SAME THING?

The promoter elements controlling temporal and seed specificity could be separate elements, or they could coincide. In the latter case, a seed-specific element could simply activate expression within the seed during particular stages of development. Since seed storage protein genes are not expressed coordinately, such seed-specific/temporal elements would have to differ between individual genes and gene classes. In the case of separate elements, the primary function of a seed-specific element would be to repress expression in non-seed tissues, while permitting activation of expression within seed tissue. Additional regulatory elements would be necessary to ensure correct temporal expression during the development of the seed.

Analyses of many seed genes have not been able to distinguish the regions controlling temporal regulation from those containing the seed-specific regulatory elements. Genes that have been analysed include those coding for the embryo-specific globulins (Higgins et al., 1988; Newbigin, 1988; Shirsat et al., 1989; Newbigin et al., 1990; Rerie et al., 1991; Fujiwara and Beachy, 1992), the seed-abundant lectins (Voelker et al., 1987; Jofuku and Goldberg, 1989) and the endosperm-specific prolamines from cereals (Marris et al., 1988, Colot et al., 1987; Robert et al., 1989; Murai and Kawagoe, 1991; Takaiwa et al., 1991b).

Detailed analyses of Pvu-β (also termed β-phaseolin), a gene coding for phaseolin in P. vulgaris, however, have distinguished discrete regions which if deleted, move the onset of expression from 16 days after flowering (DAF) to 11 DAF, without affecting seed-specific or spatial expression (Bustos et al., 1991). Thus these regions act as negative elements, delaying the onset of gene expression.

Two other elements that influence temporal expression have been identified within the 5' region of Pvu-β; -295 to -107, and (perhaps) -64 to -14. Both of these elements, however, also influence expression of phaseolin in vegetative tissues; -205/-107 represses low-level expression in stems and roots while -64/-14 promotes (in the absence of other regulating elements) basal levels of expression in seeds, stems and roots (Burow et al., 1992).
An element which contribute to intra-seed tissue and specificity of expression has been found in the Pvu-β promoter. The region between -468 to -391 of this promoter confers expression within the hypocotyl (Bustos et al., 1991).

These results indicate that temporal and seed-specific regulation of the storage protein genes can be regulated by a number of different positive and negative elements. In addition, some aspects of the temporal regulation of seed storage protein genes may be controlled by elements that are unrelated to those that are involved in seed-specific expression. Thus, the expression of seed storage protein genes may be regulated by the interactions of different temporal and seed-specific elements, and perhaps by interactions between trans-acting factors binding to these elements.

SEED-SPECIFICITY MAY BE REGULATED BY POSITIVE OR NEGATIVE ELEMENTS

At the simplest level, there are two types of regulatory element that could affect seed-specific regulation. A positive seed-specific element would stimulate transcription, presumably by interaction with a transcription factor that is abundant and active only within seed tissues. The second type, or negative element, would prevent transcription within non-seed tissues. A repressor protein binding to such a site could act in several ways to prevent transcription; direct intervention at the initiation complex, quenching of transcriptional activation by another protein bound to the 5' region through steric hindrance or induced conformational change, or competition with another factor for a binding site (see Figure 1.2). Tissue-specific changes of the activity or the abundance of the repressor could thereby permit expression within the seed, but repress it in non-seed tissue.

Seed storage protein genes may be regulated by either or both types of element or perhaps ones that combine both functions. Both types of element would be expected to be conserved within seed-specific genes. The presence of such elements in non-seed genes, however, would depend upon their mode of action.

Chimeric gene constructs, using the enhancer and basal promoter regions of the "constitutive" 35S CaMV promoter have been used to analyse gene expression directed by cis-acting elements within promoter fragments. In several cases, chimeric promoters have conferred expression typical of the corresponding intact promoter, indicating that regulation of such genes
involves both activation in the appropriate tissue/developmental stage and repression in other tissues (Ellis et al., 1987; Stougaard et al., 1987; Stockhaus et al., 1989; Keil et al., 1990; Stougaard et al., 1990). For example, ligation of nodule-specific elements from different nodule-specific genes to the 35S enhancer (-418 to -90) conferred expression within nodules and repressed expression within roots (Stougaard et al., 1987; Stougaard et al., 1990). Similarly, the anaerobic response element within the maize Adh-1 gene, coupled with the 35S enhancer, confers anaerobically-induced expression and represses expression within aerobically grown leaves (Ellis et al., 1987).

Analysis of promoter fragments of 7S genes from soybean and P. vulgaris in chimeric gene constructs, however, showed that the seed-specific elements enhance gene expression in seeds, without suppressing expression in other tissues. Nevertheless, the results indicate that both types of seed-specific element exist. For example, the region -257 to -78 of the α',β-conglycinin gene, Gma-α', when inserted at -90 within the 35S promoter, conferred 40-fold enhancement of expression in seeds, without suppressing expression in leaf, stem or root tissue (Chen et al., 1988).

Similarly, a chimeric promoter consisting of the CaMV 35S enhancer ligated to -295 to -109 of the phaseolin gene of Pvu-β conferred very high expression within seeds, and high expression in roots and leaves. Deletion of the region between -295 to -109, however, had the effect of increasing expression within the non-seed tissues two- to three-fold, whilst decreasing expression in seeds about ten-fold (Bustos et al., 1991). Similarly, 5 deletions from -227 to -107 increased expression in stems and roots (but not leaves) (Burow et al., 1992). This indicates that, in addition to enhancing gene expression in seeds, this Pvu-β segment contains one or more regulatory elements that can repress (although not completely suppress) expression in non-seed tissue. This indicates that seed-specific regulation may involve some degree of repression of expression in non-seed tissues.
1.4.3 Conserved DNA sequences in the 5' regions of seed-specific genes

PUTATIVE SEED-SPECIFIC DNA MOTIFS

The observation that seed storage protein genes from legumes and (most) cereals are expressed correctly in transgenic tobacco and petunia, indicates that the endogenous transcription factors of the hosts can recognise common cis-acting DNA sequences in the promoters. Sequence comparisons have shown potential cis-acting motifs that are conserved amongst the 5' regions of seed storage protein genes. When searching for conserved seed-specific elements, however, certain constraints regarding the mode of action of these elements should be considered.

An element that confers high-level expression in seeds could be present in a gene that is also expressed in non-seed tissues (for example, lectin genes). A negative seed-specific element, however, that actively represses expression in non-seed tissues would not be expected in genes that are highly expressed in other tissues. Nevertheless, exceptions are possible since DNA elements (and the factors that bind to them) can be influenced by both other DNA sequences and their positions in a promoter. Despite this, it is interesting to determine whether a putative seed-specific element is indeed restricted to the 5' regions of seed-specific genes and other genes that are expressed highly in seeds.

Short (6-8 bp) DNA sequences have been noted that are present in many seed-specific genes. These include the CATGCATG, CACA, ANCCCA and octanucleotide motifs in legume genes and the -300 box in cereal genes. The sections below examine how conserved these motifs actually are, and discuss the functional analysis of 5' regions containing these motifs. The relative frequency of these putative seed-specific motifs in the 5' regions of various genes of different plant promoters were assessed using a data base that contained the sequences of the 5' regions of 64 genes expressed highly in seeds and 72 non-seed genes (Table 1.2, located at the end of the chapter).

The CATGCATG element; both highly conserved and functional in seed-expressed promoters.

The CATGCATG element, defined as CATGCATG with up to one mismatch, is present (often in multiple copies) in the 5' region of many genes that are highly expressed in seeds. It is also called the RY element, because of the alternating purine and pyrimidine bases within the sequence (Dickinson et
al., 1988). This motif is present in the promoters of 19 globulin genes and 6 out of 7 lectin genes from legumes that were examined (Dickinson et al., 1988; Table 1.2). It was also observed, at lower frequencies, ranging up to 50%, in the promoters of seed-storage protein genes from other dicotyledonous plants and cereals. These rates (with one exception) are still higher than the observed frequency of the motif within non-seed genes (15%) and the estimated frequency of random occurrence of the motif within 300 bp (11%). Both patatin promoters (of the non-seed genes) contained CATGCATG elements, suggesting that this motif might also have a role in the regulation of vegetative storage protein genes (Table 1.2).

In addition to storage protein genes, the CATGCATG motif is present in legume and cereal promoters associated with germination, such as ABA- and auxin-regulated promoters and amylase gene promoters (Table 1.2 a, seed-expressed genes; "other", details in Appendix I). Curiously, it was not observed in the 5' regions of similar genes of other dicotyledons (Baumlein et al., 1992; Table 1.2). Hence the CATGCATG motif seems to be present in the majority of the seed storage protein genes from legumes, about half those from cereals, and is also prevalent in many other seed-expressed genes.

Functional analysis of the role of the CATGCATG motif has shown that this element is essential for seed-specific regulation in some promoters, for both high level expression and tissue-specificity. Deletion of 6 bp of the CATGCATG sequence within the otherwise intact promoter of the Vicia faba legumin gene LeB4 reduced expression 100-fold in seeds. As well as having a drastic effect on the level of expression in the seeds, the removal of the CATGCATG motif conferred low level expression in leaves (Baumlein et al., 1992). Similarly, the mutation of the two CATGCAT/c sequences within the promoter of the soybean α',β-conglycinin (7S) gene, Gma-α', reduced the level of gene expression in seeds and led to a low level of expression in leaves (Fujiwara and Beachy, 1992).

Internal deletion of the CATGCAT sequence from a soybean legumin gene also reduced expression by 10- to 20-fold (Lelievre et al., 1992). In addition, McCarty and Carson (cited Baumlein et al., 1992) found that CATGCATG was essential for the regulation of maize C1 gene by the y-parous gene product Vp1. C1 is a regulator gene of anthocyanin biosynthesis associated with seed maturation (Ludwig et al., 1990).
The CATGCATG element, however, is not the only element that can repress expression in non-seed tissue, as the region between -227 and -107 of *Pvu*-β prevents expression in stems and roots, although it does not contain a CATGCATG sequence (but contains an octanucleotide, an ADCCCA and two CACA motifs (Burow et al., 1992). Hence the CATGCATG motif may serve both positive and negative seed-specific functions within a promoter by inducing gene expression within seed tissue, and repressing expression in leaf tissue. Other elements may be required, however, to prevent expression in other vegetative organs, such as roots and stems.

The CATGCATG motif, although necessary, is not sufficient for high levels of expression in seeds. A 5' deletion analysis of a legumin gene from pea, *LegA1*, showed that 237 bp of the 5' flanking sequence was insufficient to permit the expression of the legumin gene in transgenic tobacco, even though it contained a CATGCATG motif (contained within the conserved legumin box). An unidentified element between -668 and -227 was required to re-establish high levels of expression (Rerie et al., 1991). Similarly, deletion analyses of other 11S and 7S genes; such as *LeB4* (Baumlein et al., 1991b) and *Gma-α'* (Chen et al., 1986) have shown that the CATGCATG sequence was not sufficient for high-level expression in seeds and that further upstream elements were necessary for detectable gene function.

In spite of its conservation and function, DNA binding assays have failed to detect nuclear proteins which bind *in vitro* to the CATGCATG element. CATGCATG elements assayed to date include those present in the 7S genes (α',β- and β,β-conglycinin: Allen et al., 1989; Lessard et al., 1991; Fujiwara and Beachy, 1992; and β-phaseolin: Bustos et al., 1989; Kawagoe and Murai, 1992), the 11S gene (*legA*: Meakin and Gatehouse, 1991) and helianthinin (Jordano et al., 1989) and lectin genes (*Le1*: Jofuku et al., 1987; *lec2*: Riggs et al., 1989). A multimerized CATGCATG oligo also failed to bind nuclear proteins *in vitro* (Riggs et al., 1989).

Nuclear proteins from oilseed rape, however, bind *in vitro* to oligonucleotides containing purine/pyrimidine stretches from the 2S napin promoter (Gustavsson et al., 1991). One of these stretches, TACACAT, does not resemble CATGCATG, but the other sequence contains a truncated form, CATGCA. It has not been demonstrated, however, by mutagenesis or footprinting whether the interaction is with the CATGCA site or other regions within the sequence. It is possible that the CATGCA motif is not the binding site, as was found to be
the case for a zein gene. The binding region of Opaque-2, a transcriptional activator, is between -307 and -284 of the 22Z-2 gene, which contains a CATGCATG motif. Mutational analysis showed however, that the core binding site is ACGT (Schmidt et al., 1992), the target site recognised by many bZIP proteins.

The inability to detect the binding of nuclear proteins to the CATGCATG motif could be due to a number of reasons. Firstly, the CATGCATG motif may not function by acting as a binding site within seed-expressed promoters. Alternatively, the nuclear protein which binds to the CATGCATG motif, may not be abundant or the complex may be unstable in vitro. Yet another hypothesis has been suggested based upon the ability of purine/pyrimidine sequences to form Z-DNA (Riggs et al., 1989; Aryan et al., 1991). It may be that the CATGCATG motif adopts an altered conformation in vivo through interaction with upstream activator proteins, which enhances the recognition or activation of transcription complexes.

Thus, the CATGCATG motif has been shown to be essential for high-level, seed-specific expression of some genes. Nevertheless, interaction with other, additional upstream elements is required and the mode of action of the motif is still unknown.

The -300 box is a prolamine-specific motif - but what does it do?

The -300 box is another highly conserved element, but unlike the CATGCATG sequence, it is virtually restricted to cereal prolamine genes, nuclear proteins have been detected binding to it, and it has no reported effect upon seed-specific expression, and may have little or no effect upon quantitative expression.

The sequence of the -300 box, TGTTAAG, is present in most prolamine genes (Table 1.2). The broader consensus sequence, TGt/a/cAAg/ag/t, is located about 300 bp upstream of the transcription start site in all zein genes of maize and is also present in the hordein genes of barley, and the gliadin and glutenin genes of wheat (Kreis et al., 1985; Forde et al., 1985; Boronat et al., 1986; Colot et al., 1987; Matzke et al., 1990; Entwistle et al., 1991; Thompson et al., 1992; and reviewed in Thompson and Larkins, 1989), kafarin genes of sorghum (DeRose et al., 1989), avenin genes of oat (Shotwell et al., 1990), and the glutenin genes of rice (Okita et al., 1989; Takaiwa et al., 1991a). In addition, the -300 box is also present in the 5' regions of other seed-expressed
genes, including pea legumin genes (Lycett et al., 1984; Lycett et al., 1985)). Only one of the non-seed genes examined contained the -300 motif, a soybean heat shock gene, (Helm et al., 1993), indicating that the motif is virtually exclusive to seed genes (Table 1.2, Appendix I).

Despite the prevalence of this motif, deletion studies indicate that the highly conserved -300 box is not essential for seed-specific expression (Thomas and Flavell, 1990; Matzke et al., 1990; Halford et al., 1989). Studies indicate that the -300 box may have a slight effect on the quantitative level of expression of some genes (Colot et al., 1987; Thomas and Flavell, 1990; Quayle and Feix, 1992), while in other genes this sequence has little or no effect (Matzke et al., 1990; Aryan et al., 1991; Thompson et al., 1990). For instance, the zein gene Z19ab1 from maize has an enhancer element between -347 and -226 which contains a -300 box. Internal deletion of 10 bp containing the -300 box sequence, however, did not affect activity of this element (Thompson et al., 1990). Conversely, a two bp mutation within the -300 box of another 19 kd zein gene, pMS1, reduced gene expression by about 30% (Quayle and Feix, 1992).

Notwithstanding the lack of evidence for a strong functional role, the -300 box is recognised in vitro by DNA-binding proteins. Endosperm-specific nuclear proteins bind to the -300 box within the pMS1 gene from maize (Maier et al., 1987; Maier et al., 1988) and that mutation of this -300 box reduces the binding of proteins (Quayle and Feix, 1992). Together these data suggest that the -300 box may have a quantitative role in some, if not all prolamine genes, and that this role is probably mediated by specifically bound nuclear proteins. It remains a puzzle, however, why this apparently non-essential element has been maintained in so many related genes.

The CACA box is present in many seed-expressed genes

The CACA element is present in the promoter regions of many cereal and dicotyledonous genes that are highly expressed in seeds, including those expressed during the later stages of seed development.

CACA elements were first noted in 11S, lectin and Kunitz trypsin inhibitor promoters from legumes (Goldberg, 1986). The core of the motif, AACACA, is present in most other seed genes from legumes and other dicots, and is also present at a high frequency in the seed-specific genes from cereals (Reeves and Okita, 1987; table 1.2). The frequency of the CACA motif in most classes
of seed-expressed genes, including those expressed at later stages of development than seed storage protein genes, is significantly greater than the expected value for random distribution of the motif (Table 1.2). This suggests that the CACA element could have a role in the activation of expression of seed genes throughout most stages of development, rather than at particular stages. In addition, the CACA element is present at lower frequencies within non-seed genes, such as light-regulated and stem-specific genes (Table 1.2, Appendix 1). Since the CACA box is moderately prevalent in non-seed genes, this motif may have a regulatory function in other genes, or the role of the motif (if any) may be modified by surrounding cis-acting elements.

CACA-like elements are also present as part of larger conserved sequences in the promoters of genes expressed in cereal grains. The AACACA sequence is in the middle of another, longer "CACA" motif (24 bp) which is present in wheat gliadin genes (Reeves and Okita, 1987) and barley hordein genes (Entwistle et al., 1991). Other CA-rich motifs include the 7-11-7 motif (CAACAAA -11N-ATGTCAA) which is present in all 19 kd, 22 kd and 27 kd zeins (Boronat et al., 1986) and the AACACA motif (AACACAACTCTAT) which is present between -78 to -65 of promoters of glutelin genes (Takaiwa et al., 1987b; Takaiwa et al., 1991a).

The CACA element is also similar to AACAAaCA (lower case indicates where the sequence differs from the core element), which is present in a gibberellin/abscisic acid response element in α-amylase genes from barley (Skriver et al., 1991; Gubler and Jacobsen, 1992) and is conserved in α-amylase genes of many species (Huang et al., 1990). Curiously, the CACA motif is also present within the H1 box, AAACACA, which modulates expression of the histone h1 genes of invertebrates and vertebrates (Dalton and Wells, 1988). It is not however, present in any of the 10 plant histone genes screened in Table 1.2 (Tabata et al., 1983; Tabata et al., 1984; Chabut et al., 1986; Peng and Wu, 1986; Chaboute et al., 1987).

It has not been established whether the CACA element has a functional role within seed-expressed genes. Lindstrom et al. (1990) claimed that the CACA element within Le-1, a soybean lectin gene, could be deleted without reducing the level of gene expression, or affecting developmental regulation. This was based, however, on visual comparisons of small sample sizes (about 3 plants/construct tested), rather than stringent quantitative data. Results of studies on a pair of lectin genes from Dolichos biflorus also cast doubt upon the
role of CACA motifs. Sequence comparison of these genes, a seed lectin and a stem- and leaf-specific lectin, showed their 5' regions were almost identical, except for a 116 bp deletion in the stem/leaf lectin gene. The seed-expressed lectin gene does not contain a CACA motif, but the stem/leaf -expressed lectin does (Harada et al., 1990). Furthermore, the seed-expressed lectin gene contains an ADCCCA motif, which is absent in the other gene, suggesting that the motif may be of functional significance. In contrast, deletion of a 76 bp region containing a CACA element from a gliadin gene caused a significant decrease in gene expression (Aryan et al., 1991).

Despite the conflicting evidence regarding the functional importance of the CACA element, DNA binding experiments indicate that nuclear proteins bind CACA-like elements. Nuclear protein CA-1 binds to a CACA element and a related sequence, AcCACA, in the 5' region of a ß-phaseolin gene from bean, Pvu-ß. Curiously, the binding protein did not recognise the second CACA element in the promoter (Kawagoe and Murai, 1992), suggesting that flanking sequences may have an important role in determining binding specificity.

CA-rich oligonucleotide sequences derived from a non-legume seed-storage protein napin (2S) gene also interact with nuclear proteins (Gustavsson et al., 1991). Two wheat endosperm nuclear factors have also been identified which interact specifically with the CACA box within a gliadin gene (cited by Aryan et al., 1991). DNA-protein binding has also been detected to CACA-like motifs within cereal genes. A region of the maize zein gene promoter, containing ACACA on the opposite strand is protected from DNase I digestion in footprinting experiments and may bind nuclear proteins. Furthermore, this region has the capacity to form a stem loop by interaction with another CACA-like region immediately upstream (Maier et al., 1987). Such secondary structure in the promoter may display or mask the DNA-binding sites and bound proteins, and so could moderate binding of nuclear proteins.

Octanucleotide box is specific to globulin and lectin genes.

The octanucleotide box GCCACc/tTC is present at about position -150 in the promoters of most 7S and 11S genes (Lycett et al., 1985; Gatehouse et al., 1986; Rerie et al., 1992) and in some lectin genes (Voelker et al., 1987; Table 1.2 and Appendix I). The octanucleotide box was first noted because it resembles part of the opposite strand of the core of the adenovirus enhancer sequence, GtCACTTC (Hearing and Shenk, 1983). It is not present in any other class of gene, from either dicotyledonous or monocotyledonous plants.
Table 1.2 and Appendix I). This motif, therefore, appears to be restricted to the 5' regions of globulin and lectin genes.

The presence of the octanucleotide motif, in conjunction with other conserved sequences, has been correlated with seed-specific expression of globulin and lectin genes. For example; region -245 to -57 of Gma-α’ (which contains both the octanucleotide and CACA motifs) conferred 40-fold enhancement of gene expression within transgenic seed, when inserted into a CaMV 35S promoter (Chen et al., 1988). A highly-expressed lectin allele from common bean, dlec2, contains two octanucleotide sequences as well as two CATGCATG motifs, while the almost identical 5' region of a poorly-expressed allele has a deletion between -354 and -240, removing these motifs (Voelker et al., 1987). A 5' deletion analysis of dlec2, showed that deletion of the region -345 to -125 eliminated gene expression, confirming that this region contains important regulatory elements (Riggs et al., 1989).

As well as being correlated with high levels of seed-specific expression, the octanucleotide box, and two AACACA motifs are also present in an element in Pvu-β that suppresses transcription in stems and roots (Burow et al., 1992)). Nuclear proteins, however, have not been observed binding in vitro to octanucleotide sequences (Kawagoe and Murai, 1992; Allen et al., 1989). Further analysis is required to determine the role of this highly-conserved motif.

Although the ADCCCA element binds nuclear proteins, it is not necessary for gene expression and is not specific to genes expressed in seeds.

Initial experiments suggested that the ADCCCA motif might play an important role in the regulation of seed-specific promoters. High levels of expression of a 7S gene from soybean, Gma-α’, correlated with the presence of 5 ADCCCA repeats in the promoter. Deletion of a region containing four copies of the repeat (located between -257 and -159) reduced the level of gene expression at least 20-fold. Further deletion of the region containing the remaining ADCCCA motif (region -159 to -69) reduced gene expression below detectable levels (Chen et al., 1986). Furthermore, Chen et al. (1988) showed that insertion of the Gma-α’ region from -245 to -57 (in either orientation) into the promoter of a constitutively-expressed reporter gene gave a 25- to 40- fold enhancement of gene expression in seeds during the mid- to late- stages of embryo development.
It has been demonstrated that a nuclear protein from immature soybean seeds, SEF3, binds exclusively at two AACCCA sites in the Gma-α' promoter. The binding was correlated with the temporal expression of Gma-α' (Allen et al., 1989). Specific mutations in the SEF3 binding sites, however, did not have a significant effect upon quantitative or seed-specific gene expression, although they did abolish binding of nuclear proteins in vitro (Fujiwara and Beachy, 1992). Thus, a role for ADCCCA in transcriptional regulation of seed genes is not yet established.

The lack of an obvious functional role for the ADCCCA motif may explain why it is not present at significantly greater than random frequency in the promoters of most other plant promoters including seed-storage protein genes. In contrast, cereal genes associated with germination, and all histone genes contained higher-than-expected frequencies of this motif (P< 10%) (Table 1.2). It is difficult, however, to assess the significance of the presence of this short motif within promoters, because it would be expected at high frequency within random DNA sequences.

**GENE-SPECIFIC ELEMENTS**

Comparisons of closely-related genes have distinguished longer conserved regions within the 5' regions, ranging up to 70 bp in length. Most of these longer, conserved regions contain one or more of the smaller conserved elements, especially the CATGCATG motif. Perhaps the arrangement of these conserved motifs and the sequences that flank them, act to determine the different temporal patterns of expression of gene-specific promoters.

**Vicilin box I**

The legume 7S genes contain a conserved element located about 100 bp upstream from the start of transcription termed 'vicilin box' or 'vicilin box I'. The motif was initially defined as a 42-bp sequence in two vicilin genes from pea (Vic B, Vic J) and one from common bean (P. vulgaris; Pvu-β; Gatehouse et al., 1986)). It contains three smaller motifs; an octanucleotide sequence, a CACA element and a CATGCATG element. A 33-bp version (reviewed in Rerie et al., 1992), also starting at the octanucleotide box, but lacking the CATGCATG element, is present in vicilin genes from pea (Psa-v, Cvc-b), field bean (Vic-1), soybean (Gma-α', Gma-β, Cgy4) and jack bean Can-1, Figure 1.4). In addition, Newbigin (1988; see also Higgins et al., 1988) has shown that the Psa-v, Cvc-b, Gma-α' and Pvu-β 7S genes share sequences extending 30 bp downstream of the CACA element. Vicilin box I is not conserved, however,
Figure 1.4. Vicilin box I sequence in 7S genes.

Derived from Gatehouse et al. 1986 and Rerie et al. 1992. Octanucleotide, CACA and CATGCATG motifs are in bold, variable bases are in lower case. Only mismatched nucleotides to the vicilin box I in *Psa-v* are shown. An overall consensus sequence is presented (33 bp).

<table>
<thead>
<tr>
<th>7S protein</th>
<th>Gene</th>
<th>Vicilin box I</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pea vicilin</td>
<td><em>Psa-v</em></td>
<td>GCCACTTCAATTTTgTACATTTCACACaCGTC</td>
<td>Higgins 1988</td>
</tr>
<tr>
<td>Pea vicilin</td>
<td><em>Vic B</em></td>
<td>c</td>
<td>Gatehouse 1986</td>
</tr>
<tr>
<td>Pea vicilin</td>
<td><em>Vic J</em></td>
<td>t</td>
<td>Gatehouse 1986</td>
</tr>
<tr>
<td>Pea convicilin</td>
<td><em>Cvc-a</em></td>
<td>t</td>
<td>Bown 1988</td>
</tr>
<tr>
<td>Pea convicilin</td>
<td><em>Cvc-b</em></td>
<td>t</td>
<td>Newbigin 1989</td>
</tr>
<tr>
<td>Field bean vicilin</td>
<td><em>Vic-1</em></td>
<td>c</td>
<td>Weschke 1987</td>
</tr>
<tr>
<td>Soybean α,β-conglycinin</td>
<td><em>Gma-α</em></td>
<td>c</td>
<td>Doyle 1986</td>
</tr>
<tr>
<td>Soybean β,β-conglycinin</td>
<td><em>Gma-β</em></td>
<td>t</td>
<td>Tierney 1987</td>
</tr>
<tr>
<td>Soybean β,β-conglycinin</td>
<td><em>CG-4</em></td>
<td>- --------------------------------------------------</td>
<td>Harada 1989</td>
</tr>
<tr>
<td><em>P. vulgaris</em> β phaseolin</td>
<td><em>Pvu-β</em></td>
<td>c</td>
<td>Slightom 1983</td>
</tr>
<tr>
<td>Jack bean canavalin</td>
<td><em>Can-1</em></td>
<td>g</td>
<td>Takei 1989</td>
</tr>
<tr>
<td>CONSENSUS</td>
<td></td>
<td>GCCACTTCAATTTTgTACATTTCACACaCGTC</td>
<td></td>
</tr>
</tbody>
</table>
within the vicilin genes from non-legumes such as cocoa (McHenry and Fritz, 1992), cotton, wheat, or any other types of genes (Table 1.2 and listed in Appendix I).

**Vicilin box II**

A second conserved element, termed vicilin box II has also been identified (Newbigin, 1988). This 20 bp motif, nTTCA TnAAT TCAAACAAAA (with up to two mismatches), is present in at least 5 of the 7S genes, located between -350 and -150 bp (Newbigin, 1988; Table 1.2 and Appendix I). The region containing vicilin box II can be deleted without significantly affecting the level of gene expression of Gma-α' (compare constructs 4-457 and 4-257; Chen et al., 1986), indicating that this sequence is not essential for expression.

**Legumin box**

The 11S globulin genes contain a highly conserved 29 bp element located at position -100, called the 'legumin box', which contains a CATGCATG motif (Baumlein et al., 1986; Gatehouse et al., 1986; Rerie et al., 1992; Figure 1.5). In spite of its conservation among legumes, the legumin box is not recognisable in the 5' flanking sequence of 11S genes from most non-legumes (Takaiwa et al., 1987a; Pang et al., 1988; Vonder Haar et al., 1988; Okita et al., 1988; Bogue et al., 1990; Schubert et al., 1990; Shotwell et al., 1990), or any other genes (Table 1.2 and Appendix I). A degenerate version of the legumin box, containing 17/28 bp sequence identity is present in the 5' region of an oilseed rape 12S gene (Ryan et al., 1989).

**-300 element**

Cereal grain storage protein genes also contain gene-specific sequences. These include various CA-rich sequences (discussed above, p35). Other motifs are extensions of the -300 box (TGTAAG). The longer -300 element, which is also referred to as the endosperm or prolamine box, has the consensus ANNTGTAAAG(G/T)(G/T)AAT NNGATGAGTCATG (bold letters are completely conserved). This sequence is present in the promoters of prolamine genes coding for wheat gliadin and glutenin, barley hordein, and maize 21 kDa zein (Forde et al., 1985; Reeves and Okita, 1987; Colot et al., 1989). The first 17 bp of the consensus are also present in maize 19 kDa zein genes (Forde et al., 1985; Roussell et al., 1988). The zein genes contain a more conserved element, CACATGTGTAAGGT, which overlaps with the
Figure 1.5 Legumin box sequence in 11S genes.

Derived from Gatehouse et al., 1988, Baumlein et al., 1986 and Rerie et al., 1992 (small errors in consensus for legJ and Gy4(R) are corrected). The CATGCATG motif is in bold, variable bases (two or more genes differing from the consensus, for any given base) are in lower case. Only mismatched nucleotides to the legumin box I in LegA2 are shown. An overall consensus sequence is presented.

<table>
<thead>
<tr>
<th>11S protein</th>
<th>Gene</th>
<th>Legumin Box</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pea legumin</td>
<td>legA2</td>
<td>T C C A T A a C C A T G C A a G a T G A A G A A T g T C</td>
<td>Rerie 1990</td>
</tr>
<tr>
<td>Pea legumin</td>
<td>legA1</td>
<td>g c g</td>
<td>Rerie 1991</td>
</tr>
<tr>
<td>Pea legumin</td>
<td>legA</td>
<td>g c g</td>
<td>Lycett 1984</td>
</tr>
<tr>
<td>Pea legumin</td>
<td>legB</td>
<td>g c g</td>
<td>Lycett 1985</td>
</tr>
<tr>
<td>Pea legumin</td>
<td>legC</td>
<td>g c g</td>
<td>Lycett 1985</td>
</tr>
<tr>
<td>Pea legumin</td>
<td>legJ</td>
<td>g t c</td>
<td>Gatehouse 1988</td>
</tr>
<tr>
<td>Soybean legumin</td>
<td>Gy1</td>
<td>g t a c</td>
<td>Sims 1989</td>
</tr>
<tr>
<td>Soybean legumin</td>
<td>Gy2</td>
<td>g t a c</td>
<td>Thanth 1989</td>
</tr>
<tr>
<td>Soybean legumin</td>
<td>GyA2B1a</td>
<td>g t a c</td>
<td>Kitamura 1990</td>
</tr>
<tr>
<td>Soybean legumin</td>
<td>Gy3</td>
<td>g t a c</td>
<td>Cho 1989</td>
</tr>
<tr>
<td>Soybean legumin</td>
<td>Gy4 (R)</td>
<td>a a g</td>
<td>Scallon 1987</td>
</tr>
<tr>
<td>Soybean legumin</td>
<td>Gy4 (F)</td>
<td>a a g</td>
<td>Xue 1992</td>
</tr>
<tr>
<td>Soybean legumin</td>
<td>Gy4 (D)</td>
<td>a a g</td>
<td>Xue 1992</td>
</tr>
<tr>
<td>Fieldbean legumin</td>
<td>LeB4</td>
<td>g t c</td>
<td>Baumlein 1986</td>
</tr>
<tr>
<td>Oilseed rape 12S</td>
<td>cruA</td>
<td>c a a t a g at g t g</td>
<td>Ryan 1989</td>
</tr>
</tbody>
</table>

CONSENSUS
endosperm box. This 15 bp element is present in 19 kDa [zein], 21-23kDa zein and kafarin (22 kDa) genes (Brown et al., 1986; Kriz et al., 1987; DeRose et al., 1989; Matzke et al., 1990; Thompson et al., 1992).

1.4.4 Binding sites of putative transcription factors

DNA-PROTEIN BINDING ASSAYS

The previous discussion concentrated on DNA sequences which are potentially important for regulating seed-storage protein gene expression based on functional analysis or sequence conservation. Conserved regions were then examined for their ability to bind nuclear proteins. A complementary experimental approach is to use DNA-protein binding assays as an initial screen for potentially important regulatory regions. This method has located DNA sequences that bind nuclear proteins. However, functional analysis indicates that not all DNA-protein binding regions are cis-acting elements regulating gene expression in the seed.

The majority of sites recognised by DNA-binding proteins are AT-rich sequences (discussed below), although other, more specific, DNA binding sites have also been identified. Surprisingly, however, apart from the CA-rich and CATGCATG-like sequences in promoters of certain seed storage protein genes (discussed above), and the G-box sequence, these other binding regions do not resemble DNA motifs that are conserved in seed-expressed genes.

G-boxes

Although by no means seed-specific, since they were first identified within light-regulated genes by Giuliano et al. (1988), G-boxes (CACGTG) are present in 5' regions of many seed-specific genes (Williams et al., 1992; and Appendix I)

A G-box-like sequence is found within 22 kDa zein promoters. This site interacts with the protein encoded by the Opaque-2 gene (Schmidt et al., 1992). This gene encodes Opaque-2 (O2), a bZIP-type protein (Schmidt et al., 1987; Motto et al., 1988; Hartings et al., 1989; Schmidt et al., 1990) characteristic of transcription factor found in other eukaryotes. Foot printing and sequence analysis indicate that O2 binds to the consensus TCCACGTAGA (Schmidt et al., 1992). Mutation of the central ACGT to ACaT inhibits binding by O2 and shows that this core sequence is necessary for DNA binding (Schmidt et al., 1992). Mutation of an O2-like target sequence within a 27 kDa zein gene
promoter so as to increase the in vitro binding affinity to the O2 protein conferred a 10-fold increase of in vitro transcription in the presence of O2 (Ueda et al., 1992).

The Opaque-2 gene is a regulatory locus in maize: Opaque-2 mutants show a severe decrease in transcription of 22kDa zein genes (reviewed in Motto et al., 1989). Together this data indicates that the transcription of the 22 kDa zein genes requires the binding of a protein factor to a G-box like sequence.

Binding of the nuclear protein termed CAN to a G-box has also been reported in a seed storage protein gene from common bean (Kawagoe and Murai, 1992). This binding site is located within the major positive element, between -295 and -228 (Burow et al., 1992). The CAN protein also binds to sequences within a negative element that represses stem/root expression, and quixotically, within an adjacent positive element that confers low level expression upon stems and roots. These other sites, CAATaTG and CACcTG, however, do not fit the G-box consensus, indicating that the preferred target sequence of the CAN protein is CANNTG (Kawagoe and Murai, 1992). Thus CAN has different sequence specificity to other G-box binding proteins such as Opaque-2 (Schmidt et al., 1992) and CG-1 (Staiger et al., 1989).

**Binding sites within 7S genes**

The following discussion shows that an abundance of nuclear proteins bind in the 5' regions of genes encoding legumin 7S proteins. However, these proteins do not appear to bind many conserved sequences, nor have the protein binding sites been shown, by functional analysis, to contain cis-acting regulatory elements.

The promoters of both 7S genes from soybean, Gma-α', and Gma-β encoding α',β- and β,β'-conglycinin, respectively, contain multiple binding sites for the Soybean Embryo Factors SEF1, SEF2 and SEF4 while Gma-α' contains two SEF3 binding sites (Allen et al., 1989; Lessard et al., 1991). The SEF1 sites, however, do not appear to contribute to gene expression. Allen et al. (1989) also suggest that SEF2 is unlikely to have a specific function, because it has low DNA sequence specificity; binding to these sites can be competed by 20 to 50 fold excess of various non-specific DNA sequences. However, SEF3 (see page 37), binds to two AACCCA sites within the 5' region of Gma-α' (Allen et al., 1989). Mutation of these sites, however, whilst preventing binding, does not affect the level or specificity of gene expression (Fujiwara and Beachy, 1992). This observation and the absence of SEF3 binding sites within the 5' region of
Gma-β (Lessard et al., 1991) indicates that this region may have little or no transcriptional role.

Similarly, the SEF4 binding sites, a/gTTTTTa/g, do not appear to be essential for gene expression. Gma-α' contains three and Gma-β contains six SEF4 binding sites (Allen et al., 1989; Lessard et al., 1991). Mutation of the proximal SEF4 site within the promoter region of Gma-α' did not significantly reduce or otherwise alter gene expression (Fujiwara and Beachy, 1992). These results do not conclusively prove that SEF4 have no function since other SEF4 sites remained in some of the mutated gene constructs in these experiments. The other mutated gene constructs contained the 35S core promoter which has elements that have been shown to synergistically interact with upstream binding sites (Benfey and Chua, 1990; Davis et al., 1990).

The 5' region of the phaseolin gene, Pvu-β is also bound by many nuclear proteins (Kawagoe and Murai, 1992). Five sets of proteins, including a probable TATA-binding factor, bind within the first 443 bp of the 5' region. Phaseolin and the β-conglycinins are all 7S genes and share 58% sequence identity over 400 bp of the promoter (Doyle et al., 1986; Chee and Slightom, 1991). Despite this, the sites in the phaseolin promoter bound by the nuclear proteins of bean are different to the sites bound by the nuclear proteins from soybean. For instance, Pvu-β's AACCCA element (an element bound by SEF3 in the bean gene) is not bound by proteins from soybean nuclei (Kawagoe and Murai, 1992). Instead, different CA-rich regions in Pvu-β are recognised by nuclear proteins from bean seeds. Two CA-1 binding sites, CACAaCACACA and cttcatCAtcAcCACAac (bases identical to each other in upper case, the CACA motif underlined) are present between -227 to -106, a region which contains the element(s) that repress(es) expression in roots and stems (Burow et al., 1992).

The AG-1 protein which is abundant in seeds, but also present in leaves (Kawagoe and Murai, 1992) binds within two regions containing negative regulatory elements; once, on the opposite strand, within the stem/root element (-227 to -106) and twice within -422 to -295, a region which contains an element that reduces the steady state level of phaseolin mRNA in seeds (Burow et al., 1992). The consensus binding site AAAAAGa/gCAA, resembles the opposite strand of the consensus SEF4 binding site.
Two nuclear proteins CAN-1 (discussed above) and C1, have been detected binding within the major positive element, located between -295 and -228. Competitive binding and footprinting experiments have defined the C1 binding site as probably within the GC-rich sequences CGCCGCGTCC (Kawagoe and Murai, 1992).

Ironically, promoter analysis of legume seed storage genes within the seed-specific regulatory regions has, so far, yielded conserved cis-acting DNA elements such as CATGCATG that have not been shown to be bound by potential trans-acting protein factors. In addition, a variety of DNA-protein binding sites have been identified, but at least some of these do not appear to act as cis-acting elements. Many quantitative regulatory regions within seed storage protein genes, however, contain AT-rich binding sites that bind similar nuclear proteins. These AT-rich sites may act as cis-acting elements that enhance gene expression, via binding of homologous nuclear factors.

Many quantitative elements have AT-rich sequences that bind abundant nuclear proteins.

AT-rich binding sites in seed-storage protein genes

Gene promoter analysis has determined the location of various elements that regulate gene expression of many seed-expressed genes. The pattern that has emerged from 5' deletion analysis is that quantitative elements are located upstream of the seed-specific and temporal regulatory elements; such that deletion of the distal upstream regions, between approximately -3000 and -300, affects the level of gene expression without affecting seed-specific expression (Table 1.1).

Quantitative elements that do not affect seed-specificity and are not essential for gene expression have been identified in the promoters of genes encoding 7S, 11S and lectin proteins (7S: Chen et al., 1986; Bustos et al., 1989; Burow et al., 1992; 11S: Nielsen et al., 1989; Shirsat et al., 1989; Bogue et al., 1990; Lelievre et al., 1992; lectin: Goldberg et al., 1989; Riggs et al., 1989). Quantitative elements have also been identified within the 5' regions of cereal prolamine genes (Thomas and Flavell, 1990; Matzke et al., 1990; Murai and Kawagoe, 1991; Takaiwa et al., 1991b).
Quantitative elements in other seed storage gene promoters have also been defined using transient gene expression (11S: Baumlein et al., 1991b; Baumlein et al., 1992; lectin: Lindstrom et al., 1990; prolamine: Aryan et al., 1991; Thompson et al., 1990; Giovinazzo et al., 1992; Brown et al., 1986). These experiments, however, did not discriminate between quantitative and seed-specific elements.

Most quantitative regions in the 5' regions of seed-storage protein genes contain AT-rich sequences. Many of these AT-rich sequences contain binding sites that are recognised by abundant DNA-binding proteins. Examples include the quantitative regions within the genes coding for phaseolin (Bustos et al., 1989; Jordano et al., 1989; Guiltinan et al., 1989), pea legumin (Shirsat et al., 1989; Meakin and Gatehouse, 1991), helianthinin (Jordano et al., 1989), lectin (Jofuku et al., 1987; Lindstrom et al., 1990), phytohaemagglutinin (Riggs et al., 1989) and 19 kDa zein (Maier et al., 1990; Grasser et al., 1990; Young et al., 1991).

In addition, AT-rich quantitative elements are present in another 19 kDa zein gene from maize and legumin genes from Vicia faba and soybean. However it has not been determined whether nuclear proteins bind within these sites (Roussell et al., 1988; Thompson et al., 1990; Baumlein et al., 1992; Lelievre et al., 1992).

AT-rich binding sites are present in other seed-expressed genes

AT-rich binding sites have also been detected within the 5' region of other seed-expressed genes, including genes coding for embryo-specific proteins from carrot (Jordano et al., 1989; Hatzopoulos et al., 1990; Goupil et al., 1992) and wheat (Pedersen et al., 1991) and maize sucrose synthase (Werr et al., 1988). In addition, a 9 bp AT-rich sequence present within an embryo-expressed maize anthocyanin Bz-1 promoter is critical for high level expression (Roth et al., 1991).

AT-rich binding sites are present in many other plant genes.

AT-rich sequences that are bound by nuclear proteins have been identified within the 5' regions of many other types of plant genes. These nodule-specific genes (Jensen et al., 1988; Metz et al., 1988; de Bruijn et al., 1989; Jacobsen et al., 1990; Forde et al., 1990); light-regulated genes (Castresana et al., 1988; Datta and Cashmore, 1989; Lam et al., 1990a; Schindler and Cashmore, 1990; Bruce et al., 1991; Pedersen et al., 1991; Manzara et al., 1991; Nantel et
al., 1990; Harrison et al., 1991; Cushman and Bohnert, 1992; Granell et al., 1992; Czarnecka et al., 1992) and other genes (Liu et al., 1990; Echevarria et al., 1992). This suggests that such AT-rich binding regions may have a common regulatory role, perhaps in the general enhancement of gene expression.

FUNCTIONAL ANALYSES OF AT-RICH BINDING SITES

Enhancement of activity in non-seed tissues by AT-rich sequences from seed genes

The proteins that bind to the AT-rich sequences within quantitative elements of legumin and phaseolin (Meakin and Gatehouse, 1991; Guiltinan et al., 1989; Jordano et al., 1989) and other AT-rich regions in the 5' regions of soybean lectin and α',β-conglycinin genes (Jofuku et al., 1987; Allen et al., 1989), are either inactive or less abundant in nuclear extracts from non-seed tissue. This would suggest that these sequences, when spliced into chimeric genes, would confer enhanced expression solely in seeds. The experimental data, however, does not support this idea. For example, a phaseolin gene, Pvu-β, contains an AT-rich binding site (-682 to -628, Bustos et al., 1989; Guiltinan et al., 1989) located in a weak quantitative region (Burow et al., 1992). Fusion of a sequence containing this AT-rich binding site, in either orientation, to a minimal 35S (-90) promoter/GUS gene construct conferred GUS expression in seeds, seedling cotyledons, hypocotyls, leaves and particularly high expression in roots (Bustos et al., 1989). Similarly, an 11S sunflower gene, HaG3A contains two almost identical AT-rich binding sites (-1527 to -75 and -739 to -75, Jordano et al., 1989). Gene constructs containing one or both of these binding sites confer expression in seeds and significant expression in non-seed organs, especially roots (Jordano et al., 1989; Bogue et al., 1990).

This loss of tissue-specificity might be an artefact because of interaction between the proteins bound to the AT-rich sites with other proteins bound to the regulatory elements in the 35S CaMV promoter. Alternatively, other sections in the 5' region of seed genes may contain regulatory elements that repress expression in non-seed tissues.
Not all AT-rich sites act as enhancers

It should not be assumed that all AT-rich binding sites are elements that enhance transcription. On the contrary, the regulatory element within -1000 to -675 of the lec2 phytohaemagglutinin gene, which contains an AT-rich binding site, is a negative regulatory element (Riggs et al., 1989). In addition, some AT-rich binding sites (Allen et al., 1989) can be deleted without affecting expression (Chen et al., 1986), suggesting that they do not act as quantitative elements.

Examples of apparently non-functional AT-rich binding sites in other plant genes include the NAT-2 site (Jacobsen et al., 1990) from a soybean leghaemoglobin gene, lbc3 (Jorgensen et al., 1991), and an AT-rich binding site from a patatin gene (Liu et al., 1990).

AT-rich binding sites have also been identified within negative elements of non-seed genes; for example, the AT-1 binding site in the tobacco cabE gene (Datta and Cashmore, 1989), and perhaps the AT-rich binding sites within the 5' region of a phosphoenolpyruvate carboxylase gene Ppc1 from the common ice plant (Cushman and Bohnert, 1992). Thus AT-rich sites and the proteins that bind to them, do not always act as activators of transcription.

Nuclear proteins bind to AT-rich sequences that may be structurally similar.

Cross competition between AT-rich sequences

Competition experiments indicate that a wide variety of AT-rich sequences can compete for binding in vitro of nuclear proteins to a particular site. Thus, the binding may depend upon structural features peculiar to AT-rich DNA, rather than upon the recognition of specific sequences. This loose specificity has been observed for proteins that bind to AT-rich sequences in the promoters of seed-specific genes. Nuclear proteins that bind to an enhancer element from an 11S sunflower gene can bind to fragments containing AT-rich binding sites from a bean 7S gene (Pvu-β), and a Lea carrot gene (DcG3; an embryo-specific gene). A region from a wound-inducible gene from sunflower (HaGX3), can also compete for binding, although less effectively. These sequences, while all containing AT-rich regions, are not closely-related, indicating relaxed sequence specificity (Jordano et al., 1989). Similarly, the upstream regions of Pvu-β, and genes encoding lectin (Le1), and Kunitz trypsin inhibitor (KTI) can compete with the AT-rich site in another lectin gene, lec2, for binding to nuclear protein from P. vulgaris cotyledons (Riggs et al., 1989).
Although all these sequences cross-compete, some degree of sequence specificity does seem to be required for binding of nuclear proteins to occur. The binding of a nuclear protein to an AT-rich site in \textit{Pvu-}β \cite{Bustos et al.,1989; Jordano \textit{et al.},1989} could not be competed by an adjacent AT-rich region. Nevertheless, substantial complementary base changes (A\textrightarrow T) within the binding region did not reduce the effectiveness of the binding site. Non-complementary (A\textrightarrow C, T\textrightarrow G) changes in the same positions were not tolerated \cite{Bustos \textit{et al.},1989}. The findings support the hypothesis that the interaction of proteins with AT-rich regions can tolerate a considerable amount of sequence variation, as long as a high AT-content is maintained.

\textit{AT-binding proteins recognise AT-rich sequences from different genes}

Since AT-rich binding sites are common to many plant genes, it has been proposed that these sites may be recognised by similar binding proteins. This is supported by the results from cross-competition binding experiments between sites from different classes of genes. AT-rich binding sites in seed-expressed genes can also compete for binding with sites from other plant genes. Thus the nuclear proteins that bind to AT-rich sequences may belong to one or more ubiquitous and evolutionarily conserved families \cite{Forde \textit{et al.},1990; Czarnecka \textit{et al.},1992}.

A protein that binds to a consensus AT-rich element within the 5' region of heat shock gene \textit{Gmhsp17.5E} from soybean can be specifically competed by oligonucleotides homologous to AT-rich binding sites present in the promoters of genes of 7S proteins(α',β-conglycinin) and lectin as well as other heat shock proteins and leghaemoglobin \cite{Czarnecka \textit{et al.},1992}. Curiously, although most sequences competed effectively, including the region between -189 to -160 of the \textit{Le1} lectin gene, a second \textit{Le1} region, between -170 to -120, which, supposedly, is bound by the same protein \cite{Jofuku \textit{et al.},1987}, was much less effective \cite{Czarnecka \textit{et al.},1992}. This indicates that although a common group of proteins may be able to bind to different AT-rich sequences, binding affinity to the different sequences varies. Competition experiments with mutated binding sites have also demonstrated that not all AT-rich sequences are equally effective as competitors, and the binding affinity to specific AT-rich sites can differ between binding proteins \cite{Forde \textit{et al.},1990; Jacobsen \textit{et al.},1990; Cushman and Bohnert, 1992}. 
Not only can different AT-rich sequences compete for binding to a range of DNA-binding proteins, albeit with varying effectiveness, but nuclear proteins that can bind to these sites have also been observed in extracts from many different organs. The nuclear proteins that bind to AT-rich sequences within the soybean lectin (Jofuku et al., 1987), and leghaemoglobin (Jensen et al., 1988) genes have been reported to be present only in the organs in which these genes are expressed, namely, developing seeds, and root nodules. When, however, multiple copies of these AT-rich binding sites were used, binding by nuclear proteins from extracts from many species and organs has been observed: nuclear proteins from pea plumules, soybean plumules and roots, corn coleoptiles, roots and florets and cauliflower florets (unpublished results cited in Czarnecka et al., 1990; Czarnecka et al., 1992). The proteins that bind to the multiple copies may not be the same proteins as those that bond to single copies, since the copies may change the affinity for the binding site or generate a cryptic sequence. Nevertheless, this indicates that proteins that bind to AT-rich sequences are present in most, if not all plant organs. The lectin binding site is also recognised by nuclear protein PNF-1 from soybean nodule extracts (Forde et al., 1990). Thus, there appears to be a common group of nuclear proteins that are present in various plant tissues, and have an affinity for certain AT-rich sequences located in the 5' regions of a wide variety of plant genes.

**Nuclear proteins that bind AT-rich sequences fall into two classes.**

Comparison of the properties of nuclear proteins that bind to AT-rich sequences, however, suggests that the proteins that bind to AT-rich sequences fall into two categories; High Mobility Group (HMG)-like proteins and AT-binding proteins (ATBPs), that lack HMG-like properties and which form low mobility complexes (Czarnecka et al., 1992). So, whilst many proteins that bind to AT-rich sequences may act as conventional trans-acting factors, mediating gene expression through protein-protein interactions, some of the proteins that bind to AT-rich sites resemble HMG proteins and thus may influence transcription via nucleosome positioning or as components of the nuclear scaffold.
High Mobility Group proteins

Transcription involves not only the recognition and binding of transcription factors to specific DNA elements, but also (and probably prior to such binding) the relaxation of chromatin structure to allow the access of RNA polymerases and transcription factors to their target sites. There are many levels of DNA compaction: within chromatin, naked DNA is condensed into superhelical DNA, and two turns of the superhelical DNA, encompassing 165 bp, are wrapped around an histone octamer, to form a nucleosome. Histone H1 interacts with both the bound DNA on the exterior of the octamer and DNA in the linker region. The 10-15 nm "beads on a string" filament can be further compacted into a coiled, solenoid structure with about six nucleosomes per turn, termed 30nm fibre. Higher levels of compaction can include the formation of loops of 5 to 200 kb, attached at intervals to the proteinaceous chromosome scaffold via AT-rich scaffold attachment regions (Slatter et al., 1991; von Kries et al., 1991; Felsenfeld, 1992; and references therein). Hence, relaxation of highly condensed DNA may be an important step in the activation of the transcription of genes.

HMG proteins appear to be preferentially associated with transcriptionally active chromatin. It is believed that these proteins may act as crude transcriptional regulators that increase the accessibility of the promoter elements within chromatin structure via competition with nucleosome assembly, formation of 30 nm fibres and/or chromosomal looping. HMG proteins are low molecular weight chromatin proteins. They are extractable in 0.35 M NaCl and soluble in 2-5% trichloroacetic acid or 5% perchloric acid (reviewed in Spiker, 1988; Bustin et al., 1990b). There are two major HMG families; HMG 14/17 and HMG 1/2, which are associated with different transcriptional roles.

Proteins of the HMG 14/17 family, like the AT-rich binding proteins from plants, have loose specificity for AT-rich sequences and may recognise and bind to general structural features within the minor groove, rather than specific bases within the major groove of DNA (Solomon et al., 1986; Elton et al., 1987; Karlson et al., 1989; Bustin et al., 1990a). Micro-injection of anti-HMG-17 antibodies into somatic cells inhibits transcription, indicating that this HMG protein is required for active chromatin structure (Einck and Bustin, 1983). HMG 14/17 proteins also have high affinity for nucleosomes, and so may have a role in modulating the structure of transcriptionally active chromatin, perhaps by interfering with nucleosome positioning and thereby disrupting the coiling of nucleosome beads into 30 nm fibres (reviewed in Bustin et al., 1990b). Alternatively, they
may act as components of the nuclear scaffold, and hence affect chromosome looping (Solomon et al., 1986).

Proteins of the HMG 1/2 family are also required for the initiation of transcription (Singh and Dixon, 1990). These HMG proteins can bend DNA, so the formation of active initiation complexes require altered DNA topology (Giese et al., 1992; and reviewed in Lilley, 1992). In addition, these non-histone proteins can repress in vitro nucleosome assembly, probably by binding to DNA in a way that inhibits assembly into core particles (Waga et al., 1989). Thus the HMG 1/2 proteins may contribute to the formation of transcriptionally active chromatin by relaxing the structure of the 30 nm fibres.

Certain nuclear proteins that bind to AT-rich sequences resemble high mobility group proteins.

HMG-like proteins that bind AT-rich sequences in seed-expressed genes, include the embryo-specific nuclear proteins from carrot that interact with an ABA-responsive gene (Goupil et al., 1992). Similar proteins have also been isolated from wheat and pea nuclei that bind in vitro to AT-rich sequences in the 5' region of a gene member of the embryo maturation (Em) protein family from wheat (Pedersen et al., 1991). HMG-like nuclear proteins from maize endosperm have also been reported to bind to the CAAT and TATA boxes of a zein gene (Grasser et al., 1990; Maier et al., 1990). These experiments were done without the addition of non-specific competitor, and so the binding may be due to non-specific HMG-DNA interactions (Pedersen et al., 1991), between the negatively charged DNA and positively charged residues in the HMG-like proteins.

Proteins with HMG-like properties that bind to AT-rich sequences include the heat-resistant proteins, termed NAT1 and LAT1, from soybean nodules and leaves, respectively (Jacobsen et al., 1990); the 23 kDa and 32 kDa binding proteins from heat-shocked soybean plumules (Czarnecka et al., 1992); and a developmentally-controlled, nodule-specific DNA-binding protein from Sesbania rostrata (de Bruijn et al., 1989).

Plant genes encoding HMG-like proteins have been cloned from cDNA libraries of seed mRNA from soybean (Laux and Goldberg, 1991; Laux et al., 1991), and maize (Grasser and Feix, 1991). They contain DNA-binding domains that are present in mammalian HMG proteins such as the HMG-box (Laux and Goldberg, 1991; Grasser and Feix, 1991) and the AT-hook, which is a dAdT
binding domain (Grasser and Feix, 1991). This suggests that HMG proteins from plants may have similar functions to those found in other eukaryotes.

**AT-binding proteins**

Not all proteins that bind to AT-rich sequences within seed-expressed genes appear to be HMG-like. These other proteins have been classified as AT-binding proteins (ATBPs). Based upon mobility of the DNA-protein complexes, heat stability and TCA solubility (Czarnecka et al., 1992), the seed-related ATBPs include the 60 kDa binding protein from soybean seed (Jofuku et al., 1987) and complex II proteins from sunflower seed and seedlings (Jordano et al., 1989).

Similarly, proteins that lack HMG-like characteristics bind to AT-rich sequences in other plant genes. These include the 46 kDa and 69 kDa from soybean plumules (Czarnecka et al., 1992), NAT2 from soybean nodules and leaves (Jacobsen et al., 1990), and PNF-1 from *P. vulgaris* nodules (Forde et al., 1990), AT-1 of pea (Datta and Cashmore, 1989) and NER from radish (Echevarria et al., 1992).

Whether these two groups of nuclear proteins that specifically bind to AT-rich sequences, the HMG-like and the ATBP proteins, have similar roles and functions in seed and other plant gene regulation, remains to be determined.
1.5 Aim of thesis

By 1989, many genes coding for seed storage protein genes from legumes had been sequenced. Comparisons of the promoter regions from these different seed-storage protein genes identified several conserved DNA sequences, such as the CATGCATG element and vicilin box I, that were postulated to contain regulatory elements associated with gene expression (legume genes: Lycett et al., 1985; Baumlein et al., 1986; Chen et al., 1986; Gatehouse et al., 1986; Goldberg, 1986; Dickinson et al., 1988; cereal genes: Forde et al., 1985; Kreis et al., 1985; Brown et al., 1986; Takaiwa et al., 1987b).

It had also been demonstrated that many of the conserved DNA sequences in legumes were contained in a short region of the promoter of the \( \alpha', \beta \)-conglycinin seed storage protein gene from soybean, and that this portion of the promoter conferred seed-specific expression in transgenic petunia (Chen et al., 1986). In contrast with these results, however, initial promoter analysis of pea genes coding for vicilin (\( Psa-v \), Newbigin, 1988) and legumin (\( Leg A1 \), Newbigin, 1988), indicated that the conserved sequences within these genes were not sufficient for expression in a transgenic host and that additional regulatory sequences were required.

Since the presence of the conserved elements, vicilin box II, vicilin box I (containing an octanucleotide box and a CACA box) and a CATGCATG element in the 5' region of \( Psa-v \) did not result in detectable levels of gene expression, it is possible that these conserved DNA sequences do not contain conserved regulatory elements, and that the regulatory elements controlling the high-level, seed-specific expression of \( Psa-v \) are located upstream of these elements. It is more probable, however, that some (or all) of these conserved sequences contain regulatory elements that control or modify seed-specific and temporally regulated expression, but that interactions between the conserved DNA elements with other, upstream, elements, are required for effective levels of transcription. In the latter case, then such interactions might be via DNA-DNA interactions (through the formation of secondary DNA structures) or interactions between trans-acting proteins bound to the cis-acting elements.
The aim of this research was to investigate how this vicilin gene is regulated; that is

- to locate the principal DNA regions that regulate the quantitative, temporal and seed-specific expression of the gene,
- to determine whether DNA-binding proteins (that might act as transcription factors) specifically recognise and bind to DNA sequences within these regions, and
- to investigate the role of the conserved elements, and the DNA-protein binding sites in the regulation of gene expression.

Gene constructs were made with 5' deletions in the 5' region of \( Psa-v \). These were used to determine the location of the regulatory DNA elements within the 5' region. The expression of these gene constructs were analysed in stably transformed tobacco. Gel-shift and competition assays were used to locate DNA sequences that were specifically bound by nuclear proteins from developing pea cotyledons, and to investigate the organ-specific and temporal regulation of the binding activity/abundance of these nuclear proteins.

Once the position of the seed-specific/temporal regulatory element(s) and binding sites of nuclear proteins within this region had been established, further gene constructs were made. These contained 5' deletions, mutations and duplications within this region of \( Psa-v \), ligated to a GUS reporter gene. Analysis of the expression of these constructs, in transiently transformed pea organs were used to analyse the effect of loss, mutation and duplication of conserved sequences and binding sites, upon the seed-specific, temporal and quantitative expression.
Table 1.1
Promoter analysis of seed protein genes

When promoter analysis used the coding region of the seed gene, rather than a reporter gene, then regulatory elements were assumed to be upstream of the start of transcription site (+1*) or translation site (+1^). Seed-specificity, #, implies that the region has been shown to contain seed-specific (or seed-enhancing) regulatory elements. Deletion of a region which results in a loss of detectable activity pinpoints a region that contains important regulatory elements, but not necessarily seed-specific ones, and so sequences containing such DNA region are shown in the quantitative elements column. The location of elements that moderate temporal expression, or reduce gene expression (negative elements) are listed. The location of regions that are bound by DNA-binding proteins, and whether the region was AT-rich, consensus binding sequences, and/or the name of the binding protein are listed in the binding and comments columns.
### Embryo-Specific Genes

#### Globulins

<table>
<thead>
<tr>
<th>7S Proteins</th>
<th>Gene</th>
<th>Organism</th>
<th>Seed Specificity</th>
<th>Temporal</th>
<th>Positive</th>
<th>Negative</th>
<th>Binding</th>
<th>Authors</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vicilin</td>
<td>Psa-v</td>
<td>Pea</td>
<td>-2500 to +1*</td>
<td>-2500 to +1*</td>
<td></td>
<td></td>
<td></td>
<td>Higgins 1988</td>
<td>Vicilin was correctly targeted to protein bodies, predominantly in embryos, but also in endosperm, of transgenic tobacco seed.</td>
</tr>
<tr>
<td>Vicilin</td>
<td>Psa-v</td>
<td>Pea</td>
<td>nd</td>
<td>-2500 to -49</td>
<td>-2500 to -227</td>
<td></td>
<td></td>
<td>Newbigin 1989</td>
<td></td>
</tr>
<tr>
<td>Vicilin</td>
<td>Psa-v</td>
<td>Pea</td>
<td>-306 to +1</td>
<td>-306 to +1*</td>
<td>-773 to -425 (++)</td>
<td>-230 to -194 (general)</td>
<td>-512 to -486; AT-2</td>
<td>Quigg 1993</td>
<td>Thesis, Chapters 2, 3 and 4 Number of pluses indicates relative contribution of positive elements. AT-2 consensus binding sites: GATAgATTAAT TEMP-1 consensus sites (putative) TTTTcTTT Tgacacat The three regions required for TEMP-1 binding contain vicilin box II and the octanucleotide and CACA motifs within vicilin box I, respectively.</td>
</tr>
<tr>
<td>Convicilin</td>
<td>Cvc-b</td>
<td>Pea</td>
<td>-1000 to +1*</td>
<td>-1000 to +1*</td>
<td></td>
<td></td>
<td></td>
<td>Newbigin 1990</td>
<td></td>
</tr>
<tr>
<td>α', β-conglycinin</td>
<td>Gma-α'</td>
<td>Soybean</td>
<td>-159 to +1*</td>
<td>nd</td>
<td>-257 to -208</td>
<td>-61 to -4</td>
<td>TEMP-2 and TEMP-3</td>
<td>Chen 1986</td>
<td>Gene is also referred to as CG-1 in papers from Goldberg's lab -257 to -208: 1 x ADCCCA -207 to -160: 3 x ADCCCA -159 to -70: 1 x ADCCCA Chimeric construct with 3SS promoter plus region -257/78 enhanced expression in seed but did not repress expression in non-seed tissues.</td>
</tr>
<tr>
<td>α', β-conglycinin</td>
<td>Gma-α'</td>
<td>Soybean</td>
<td>-257 to -78</td>
<td>-257 to -78</td>
<td></td>
<td></td>
<td></td>
<td>Chen 1988</td>
<td></td>
</tr>
<tr>
<td>α', β-conglycinin</td>
<td>Gma-α'</td>
<td>Soybean</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Allen 1989</td>
<td>SEF3 binding activity was temporally regulated and not present in (leaf + stem) nuclear extract. Both AACCACA sites were required for binding. SEF-4 consensus binding sequence was a/gTTTgg.</td>
</tr>
<tr>
<td>protein</td>
<td>gene</td>
<td>organism</td>
<td>seed specificity *</td>
<td>temporal *</td>
<td>positive</td>
<td>negative</td>
<td>binding</td>
<td>authors</td>
<td>comments</td>
</tr>
<tr>
<td>-----------------</td>
<td>---------</td>
<td>----------</td>
<td>--------------------</td>
<td>------------</td>
<td>---------------------------------</td>
<td>---------------------------------</td>
<td>--------------------------</td>
<td>--------------------------</td>
<td>-------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>α',β-conglycinin</td>
<td>Gma-α'</td>
<td>soybean</td>
<td>-255 to -57 (deletion of -245 CATGCAC reduced seed expression and led to expression in the leaves)</td>
<td>-245 CATGCAC mid/late</td>
<td>-245 CATGCAC, -57 CATGCAT NOT SEF3 binding sites NOT SEF4 binding site (-115/-105)</td>
<td>.</td>
<td>SEF3: AACCCA...AACCCA SEF4: ATTTTG</td>
<td>Fujiwara 1992</td>
<td>Bases in bold were mutated to Gs, these mutations disrupted binding of the respective nuclear proteins. Mutation of SEF3 (AACCCA) and SEF4 (attttt) binding sites did not affect expression</td>
</tr>
<tr>
<td>α',β-conglycinin</td>
<td>Gma-α'</td>
<td>soybean</td>
<td>-82 to +13 (enhancing) -828 to -83 (seed-specific when linked to 35S promoter -90))</td>
<td>nd</td>
<td>-245 CATGCAC, -57 CATGCAT</td>
<td></td>
<td>.</td>
<td>Chamberland 1992</td>
<td>Mutation of both RY sites decreased expression to 10%, no detectable expression in leaves, unlike Fujiwara et al (1992).</td>
</tr>
<tr>
<td>β,β-conglycinin</td>
<td>Gma-β</td>
<td>soybean</td>
<td>+1000 to +1*</td>
<td>nd</td>
<td></td>
<td></td>
<td>.</td>
<td>Bray 1987</td>
<td>Clone pJBC794</td>
</tr>
<tr>
<td>β,β-conglycinin</td>
<td>Gma-β</td>
<td>soybean</td>
<td>+1000 to +1*</td>
<td>nd</td>
<td></td>
<td></td>
<td>.</td>
<td>Lessard 1991</td>
<td>SEF-2, -3, -4 binding were developmentally regulated.</td>
</tr>
<tr>
<td>β,β-conglycinin</td>
<td>CG 4</td>
<td>soybean</td>
<td>+1045 to +1*</td>
<td>+1045 to +1*</td>
<td></td>
<td></td>
<td>.</td>
<td>Barker 1988</td>
<td>Embryo-specific expression.</td>
</tr>
<tr>
<td>β-phaseolin</td>
<td>Pvu-β</td>
<td>Phaseolus vulgaris</td>
<td>-863 to +1*</td>
<td>-863 to +1*</td>
<td></td>
<td></td>
<td>.</td>
<td>Sengupta-Gopalan 1985</td>
<td>Embryo-specific expression. Pvu-β is also termed β-phaseolin</td>
</tr>
<tr>
<td>β-phaseolin</td>
<td>Pvu-β</td>
<td>Phaseolus vulgaris</td>
<td>-795 to +20</td>
<td>nd</td>
<td>-682 to -628</td>
<td>-670/-620, AT-rich</td>
<td>.</td>
<td>Bustos 1989</td>
<td>55 bp element is AT-rich sequence that bind nuclear protein, the element has two inverted repeats, ttaATTTaag and attATTTaat. The 55 bp region containing the binding site gives expression in roots and other tissues.</td>
</tr>
<tr>
<td>β-phaseolin</td>
<td>Pvu-β</td>
<td>Phaseolus vulgaris</td>
<td>-783 to -381, AT-rich</td>
<td></td>
<td></td>
<td></td>
<td>.</td>
<td>Guiltinan 1989</td>
<td>58 kDa protein from carrot somatic embryos binds to phaseolin fragment. Protein developmentally controlled and not observed in plantlet or leaf extract.</td>
</tr>
<tr>
<td>β-phaseolin</td>
<td>Pvu-β</td>
<td>Phaseolus vulgaris</td>
<td>-671 to -627, AT-rich</td>
<td></td>
<td></td>
<td></td>
<td>.</td>
<td>Jordano 1989</td>
<td>Binding proteins also present in seedling hypocotyl extract. Also see helianthinin data, below.</td>
</tr>
<tr>
<td>GLOBULINS</td>
<td>7S proteins</td>
<td>gene</td>
<td>organism</td>
<td>seed specificity #</td>
<td>temporal</td>
<td>positive</td>
<td>negative</td>
<td>binding</td>
<td>authors</td>
</tr>
<tr>
<td>-----------</td>
<td>-------------</td>
<td>------</td>
<td>----------</td>
<td>-------------------</td>
<td>---------</td>
<td>----------</td>
<td>----------</td>
<td>---------</td>
<td>---------</td>
</tr>
<tr>
<td>β-phaseolin</td>
<td>Pvu-β</td>
<td>Phaseolus vulgaris</td>
<td>-295 to -109</td>
<td>-518 to -418 (hypocotyl expression) -295 to -109</td>
<td>-468 to -391</td>
<td>-518 / -418 (leaf/root)</td>
<td>-391 / -295</td>
<td>Bustos 1991</td>
<td>-295 to -109: expression was located within the cotyledon and shoot meristem of the embryonic axis. This region conferred seed-specific expression when linked to the 3SS-TATA sequence. It also reduced expression in leaves and roots driven by the 3SS enhancer, but did not repress it totally.</td>
</tr>
<tr>
<td>β-phaseolin</td>
<td>Pvu-β</td>
<td>Phaseolus vulgaris</td>
<td>-295 / -107 (late stage) -64 to -14 (early expression; in seed, also low expression in roots and stems)</td>
<td>-782 to -423 (minor) -295 to -228 (major) -105 / -65 (permits equal expression in shoots, roots and seeds) 64 to -14</td>
<td>-422 / -295 (minor) -227 / -107 (stems + roots)</td>
<td>Burow 1992</td>
<td>Non-specific region, -105 to -65, contains 3' part of vic I, but lacks (complete) RY, octanucleotide and CACA boxes.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>β-phaseolin</td>
<td>Pvu-β</td>
<td>Phaseolus vulgaris</td>
<td>-295 / -107 plants lacking this element had phaseolin in stem cortices and secondary roots (but not in leaves)</td>
<td>-356 to -347 AG-1 -289 to -277 GC-rich -248 to -243 CANNTG -201 to -192 CA-rich -191 to -182 AG-1 -175 to -160 CA-rich -163 to -158 CANNTG -100 to -95 CANNTG -43 to -272 x TATA</td>
<td>Kawagoe 1992</td>
<td>No binding to CATGCATG element or to the AACCCA site. Binding to distal CACA element, but not proximal one. AG-1 binds AAAAAGGAACA GC binds approximately GATCCCGCGTCC CA-1 binds to CACAAACACA and to ccTCCATCACAACACAac</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11S proteins gene</td>
<td>gene</td>
<td>organisms</td>
<td>seed specificity*</td>
<td>temporal</td>
<td>positive</td>
<td>negative</td>
<td>binding</td>
<td>authors</td>
<td>comments</td>
</tr>
<tr>
<td>-------------------</td>
<td>------</td>
<td>-----------</td>
<td>------------------</td>
<td>----------</td>
<td>----------</td>
<td>----------</td>
<td>---------</td>
<td>---------</td>
<td>----------</td>
</tr>
<tr>
<td>glycinin G1 (A1aB1b)</td>
<td>Gy 1</td>
<td>soybean</td>
<td>-66 to +1*</td>
<td>nd</td>
<td>-454 to -66</td>
<td>-427, centred at CACA motif</td>
<td>Nielsen 1989</td>
<td></td>
<td></td>
</tr>
<tr>
<td>legumin</td>
<td>LeB4</td>
<td>Vicia faba</td>
<td>-2400 to +1</td>
<td>nd</td>
<td>-2400 to -1200 (minor)</td>
<td>AT-rich sequences in -2300 to -400 internal deletion of CATGCAT reduces expression 10-20 fold</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>legumin</td>
<td>LeB4</td>
<td>Vicia faba</td>
<td>-95 CATGCATG internal deletion of motif reduced activity in seeds and permitted low expression in leaves.</td>
<td>nd</td>
<td>-566 to -471, -279 to -232, -151 to -68, -45 to -14</td>
<td>Baumlein 1992</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>legumin</td>
<td>Leg A</td>
<td>pea</td>
<td>-1239 to +1*</td>
<td>nd</td>
<td>nd</td>
<td>Presence of legumin box (at -100) was not sufficient for high level expression. Also, deletion of this element did not further reduce expression levels.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>legumin</td>
<td>Leg A</td>
<td>pea</td>
<td>-549 to +1*</td>
<td>-549 to +1*</td>
<td>-833 to -549, -549 to -97</td>
<td>Fragments -156 to -77 and -1200 to -193 enhanced activity in seeds, but more so in leaves. Motifs within elements: -566 to -471: &quot;GC&quot; box + AT-rich (Jofuku AT box) -279 to -232: 11 bp purines -151 to -68: leg box + 2X CACA-like boxes -45 to -14: TATA box</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>legumin</td>
<td>Leg A</td>
<td>pea</td>
<td>nd</td>
<td>nd</td>
<td>-549 to -124</td>
<td>Shirsat 1989</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>legumin</td>
<td>Leg A</td>
<td>pea</td>
<td>-549 to +1*</td>
<td>-549 to +1*</td>
<td>-833 to -582, AT-rich -549 to -316, AT-rich</td>
<td>Legumin box downstream of -124.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>legumin</td>
<td>Leg A1</td>
<td>pea</td>
<td>-668 to +1*</td>
<td>-668 to +1*</td>
<td>-668 to -237</td>
<td>Meakin 1991</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>legumin</td>
<td>Leg A1</td>
<td>pea</td>
<td></td>
<td></td>
<td></td>
<td>Rene 1989</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>legumin</td>
<td>Leg A1</td>
<td>pea</td>
<td></td>
<td></td>
<td></td>
<td>Rene 1991</td>
<td>Legumin box downstream of -237</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11S proteins</td>
<td>gene</td>
<td>organisms</td>
<td>seed specificity*</td>
<td>temporal</td>
<td>positive</td>
<td>negative</td>
<td>binding</td>
<td>authors</td>
<td>comments</td>
</tr>
<tr>
<td>--------------</td>
<td>------</td>
<td>-----------</td>
<td>-------------------</td>
<td>----------</td>
<td>----------</td>
<td>----------</td>
<td>--------</td>
<td>---------</td>
<td>----------</td>
</tr>
<tr>
<td>helianthinin</td>
<td>HaG3D</td>
<td>sunflower</td>
<td>-725 to -322 fragment conferred high expression in seeds and low expression in leaves, (and significant expression in roots Bogue et al 1990, see below)</td>
<td>nd</td>
<td>nd</td>
<td>nd</td>
<td>-705 to -654, AT-rich</td>
<td>Jordano 1989</td>
<td>Binding to this site was competed by AT-rich upstream sequences from HaG3A and an embryo-specific gene from carrot, DcG3, and phaseolin. Nuclear proteins that formed complex II were present only in nuclear extract from embryos, but complex I proteins were also present in seedling hypocotyl extracts.</td>
</tr>
<tr>
<td>helianthinin</td>
<td>HaG3A</td>
<td>sunflower</td>
<td>nd</td>
<td>nd</td>
<td>nd</td>
<td>nd</td>
<td>-1463 to -1514, AT-rich</td>
<td>Jordano 1989</td>
<td>Only 3 bp differences between the 3 HaG3A/D sites, but phaseolin very different: structural similarities?</td>
</tr>
<tr>
<td>helianthinin</td>
<td>HaG3A</td>
<td>sunflower</td>
<td>-2376 to +24</td>
<td>-2376 to +24</td>
<td>-1527 to -75, -739 to -75</td>
<td>nd</td>
<td>Bogue 1990</td>
<td>Fragments -1527 to -75 and -739 to -75 (and -725 to -322 HaG3D) conferred expression in different tissues, especially roots.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>2S proteins</th>
<th>gene</th>
<th>organisms</th>
<th>seed specificity*</th>
<th>temporal</th>
<th>positive</th>
<th>negative</th>
<th>binding</th>
<th>authors</th>
<th>comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>napin</td>
<td>p3NA</td>
<td>rapeseed</td>
<td>-300 to +1*</td>
<td>nd</td>
<td>nd</td>
<td>binding to oligos representing sequences; -192/-174, -107/-83, (CA-rich repeats), -143/-132, (CA-rich); -53/-46,(2 x CATGCA) -32/-26, TATA box.</td>
<td>Gustavsson 1991 Ericson 1991</td>
<td>Binding was not temporally regulated or tissue-specific; DNA binding proteins were present in seed nuclear extracts before and after napin expression had commenced and were present in leaf nuclear extracts.</td>
<td></td>
</tr>
<tr>
<td>napin</td>
<td>nap A</td>
<td>rapeseed</td>
<td>nd</td>
<td>nd</td>
<td>nd</td>
<td>nd</td>
<td>binding to oligos representing sequences; -192/-174, -107/-83, (CA-rich repeats), -143/-132, (CA-rich); -53/-46,(2 x CATGCA) -32/-26, TATA box.</td>
<td>Gustavsson 1991 Ericson 1991</td>
<td>Binding was not temporally regulated or tissue-specific; DNA binding proteins were present in seed nuclear extracts before and after napin expression had commenced and were present in leaf nuclear extracts.</td>
</tr>
<tr>
<td>albumin</td>
<td>AT2S1</td>
<td>Arabidopsis</td>
<td>-1800 to +1*</td>
<td>-1800 to +1*</td>
<td>nd</td>
<td>nd</td>
<td>De Clercq 1990</td>
<td>25 albumin was correctly targeted to protein bodies in endosperm and embryo cells of transgenic tobacco.</td>
<td></td>
</tr>
<tr>
<td>LECTINS</td>
<td>gene</td>
<td>organisms</td>
<td>seed specificity*</td>
<td>temporal</td>
<td>positive</td>
<td>negative</td>
<td>binding</td>
<td>authors</td>
<td>comments</td>
</tr>
<tr>
<td>-------------------------------</td>
<td>------</td>
<td>-----------</td>
<td>-------------------</td>
<td>----------</td>
<td>----------------</td>
<td>----------</td>
<td>--------------------------------------</td>
<td>-------------------------------</td>
<td>--------------------------------------------------------------------------</td>
</tr>
<tr>
<td>seed (+ leaf + root) lectin</td>
<td>Le 1</td>
<td>soybean</td>
<td>-77 to +1 *</td>
<td></td>
<td>-1250 to -500</td>
<td></td>
<td>AT-rich binding sites: -184 to -173, -165 to -126: consensus ATTAT/AAT</td>
<td>Okamuro 1986, and cited in Goldberg 1989</td>
<td>Developmentally regulated and seed-specific 60 kDa binding protein; also binds to embryo-expressed Kunitz trypsin inhibitor, but not to nodule-specific leghaemoglobin.</td>
</tr>
<tr>
<td>seed (+ leaf + root) lectin</td>
<td>Le 1</td>
<td>soybean</td>
<td>nd</td>
<td>-209 to +6</td>
<td>-1700 to -338 AT-rich -209 to -190 -190 to +6</td>
<td></td>
<td></td>
<td>Jofuku 1986</td>
<td>CACA box, between -243/-208, was deleted without affecting expression. -206/-190 deletion appeared to repress expression in early developmental stages.</td>
</tr>
<tr>
<td>seed (+ leaf + root) lectin</td>
<td>Le 1</td>
<td>soybean</td>
<td>nd</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Lindstrom 1990</td>
<td>Seed-expressed gene has ADCCCA, but not CACA, vice versa for other gene.</td>
</tr>
<tr>
<td>seed lectin stem + leaf lectin</td>
<td>SL</td>
<td>Dilichos</td>
<td>116 bp deletion in DB58 5' end at -215/-100 in SL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Harada 1990</td>
<td>Correct glycosylation, Gogi-processing and targeting to protein bodies in embryo and endosperm in transgenic tobacco seed.</td>
</tr>
<tr>
<td>phytohaemagglutinin</td>
<td>dlec2</td>
<td>Phaseolus vulgaris</td>
<td>-1200 to +1 *</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Sturm 1985</td>
<td>Two octanucleotide boxes and most of two 63 bp repeats containing CATGCATG motifs are absent in Pdlec2</td>
</tr>
<tr>
<td>phytohaemagglutinin</td>
<td>dlec2</td>
<td>Phaseolus vulgaris</td>
<td>-1188 to +1 *</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Voelker 1986</td>
<td>Two octanucleotide boxes and most of two 63 bp repeats containing CATGCATG motifs are absent in Pdlec2</td>
</tr>
<tr>
<td>Highly expressed</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Voelker 1987</td>
<td>Two octanucleotide boxes and most of two 63 bp repeats containing CATGCATG motifs are absent in Pdlec2</td>
</tr>
<tr>
<td>Weakly expressed</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>phytohaemagglutinin</td>
<td>lec2</td>
<td>Phaseolus vulgaris</td>
<td>-345 to +1 *</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Riggs 1989</td>
<td>Binding was NOT detected to oligo with multiple CATGCATG sites, nor to AT-rich sites within lec1 allele.</td>
</tr>
<tr>
<td>(= dlec 2)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OTHER</td>
<td>gene</td>
<td>organisms</td>
<td>seed specificity*</td>
<td>temporal</td>
<td>positive</td>
<td>negative</td>
<td>binding</td>
<td>authors</td>
<td>comments</td>
</tr>
<tr>
<td>---------------------</td>
<td>------------</td>
<td>-----------</td>
<td>-------------------</td>
<td>----------------</td>
<td>-------------</td>
<td>-------------</td>
<td>--------------------------------</td>
<td>--------------</td>
<td>--------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Kunitz trypsin</td>
<td>KT11</td>
<td>soybean</td>
<td>-1600 to +1*</td>
<td>-1600 to +1*</td>
<td>-335 to +1*</td>
<td>-335 to +1*</td>
<td>-3000 to +1*</td>
<td>Jofuku 1989</td>
<td>This gene resembles Lata Embryogenesis Abundant genes, and is regulated by ABA. Fragment -613 to -600 binds nuclear proteins from both embryo and callus nuclear extracts, other fragments bind only embryo-specific proteins.</td>
</tr>
<tr>
<td>seed + leaf + stem</td>
<td>KT12</td>
<td>soybean</td>
<td>-335 to +1*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>seed + leaf + stem</td>
<td>KT13</td>
<td>soybean</td>
<td>-3000 to +1*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>seed + leaf + stem</td>
<td>KT14</td>
<td>soybean</td>
<td>-4000 to +1*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>lipid body membrane</td>
<td>DC 59</td>
<td>carrot</td>
<td>nd</td>
<td>nd</td>
<td></td>
<td></td>
<td>-655 to -643 AT-rich</td>
<td>Hatzopolous 1990</td>
<td></td>
</tr>
<tr>
<td>(embryo-specific)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-613 to -600 AT-rich</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-474 to -457 AT-rich</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-446 to -425 AT-rich</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(also expressed in roots)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>late embryogenesis</td>
<td>DC8</td>
<td>carrot</td>
<td>-505 to +55</td>
<td>nd</td>
<td>-1528 to -505</td>
<td>-505 to -301 AT-rich</td>
<td>-32 to +78 TATA</td>
<td>Goupi 1992</td>
<td>Binding proteins not present in nuclear extracts from leaves or roots.</td>
</tr>
<tr>
<td>abundant protein</td>
<td></td>
<td></td>
<td>(expressed in all seed tissues)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Seed acyl carrier</td>
<td>ACPO5</td>
<td>Brassica</td>
<td>-1400 to +1</td>
<td>-1400 to +1</td>
<td></td>
<td></td>
<td></td>
<td>de Silva 1992</td>
<td>The G-box sequence, CACGTG at -143, is part of the ABA-responsive element CACGTGAGC.</td>
</tr>
<tr>
<td>protein</td>
<td></td>
<td>napus</td>
<td>(low expression in leaves)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Endosperm-specific Genes

<table>
<thead>
<tr>
<th>Prolamins</th>
<th>Gene</th>
<th>Organism</th>
<th>Seed Specificity*</th>
<th>Temporal</th>
<th>Positive</th>
<th>Negative</th>
<th>Binding</th>
<th>Authors</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>15 kDa zein</td>
<td>Z151</td>
<td>maize</td>
<td>nd</td>
<td>nd</td>
<td>-200 to +1*</td>
<td>-67 to -46 : contains 2 x GCAAC</td>
<td>Goldsborough 1986</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15 kDa zein</td>
<td>Z15A</td>
<td>maize</td>
<td>nd</td>
<td>nd</td>
<td>--</td>
<td>--</td>
<td>So and Larkins 1991</td>
<td>28 kDa protein beta-1 was not present in seedling nuclear extracts (but no inhibition of endosperm proteins by seedling extract). Another protein, was present in both endosperm and seedling extracts, bound within -129 to +52 of Z19ab1.</td>
<td></td>
</tr>
<tr>
<td>19 kDa zein</td>
<td>pMS1</td>
<td>maize</td>
<td>nd</td>
<td>nd</td>
<td>--1300 to +1*</td>
<td>--</td>
<td>Langridge 1984</td>
<td>Transcription of zein in yeast cells.</td>
<td></td>
</tr>
<tr>
<td>19 kDa zein</td>
<td>pMS1</td>
<td>maize</td>
<td>nd</td>
<td>nd</td>
<td>-1555 to -965</td>
<td>-339/-318 (-300 core)</td>
<td>Brown 1986</td>
<td></td>
<td></td>
</tr>
<tr>
<td>19 kDa zein</td>
<td>pMS1</td>
<td>maize</td>
<td>nd</td>
<td>nd</td>
<td>-770 to -10</td>
<td>-278 to -285</td>
<td>Maier 1987</td>
<td>22 base binding site include 14/15 bp conserved within zein genes (including -300 box), binding region has the potential to form a stem loop which involves an adjacent CACA motif.</td>
<td></td>
</tr>
<tr>
<td>19 kDa zein</td>
<td>pMS1</td>
<td>maize</td>
<td>nd</td>
<td>nd</td>
<td>b -1207 to -1187 c P1 promoter and SV40 core f -300 element (15 bp) g P2 promoter</td>
<td>Maier 1988</td>
<td>DNA fragments c and g were bound by nuclear proteins from both seedling and endosperm extracts, whilst fragments b and f were bound only by endosperm proteins.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>19 kDa zein</td>
<td>pMS1</td>
<td>maize</td>
<td>-1558 to -8</td>
<td>nd</td>
<td>nd</td>
<td></td>
<td>Schwall 1988</td>
<td>Zain promoter was active in endosperm protoplasts but not in leaf protoplasts.</td>
<td></td>
</tr>
<tr>
<td>19 kDa zein</td>
<td>pMS1</td>
<td>maize</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Grasser 1990 Maier 1990</td>
<td>HMG proteins bound to the AT-rich sites.</td>
<td></td>
</tr>
<tr>
<td>PROLAMINES</td>
<td>gene</td>
<td>organisms</td>
<td>seed specificity*</td>
<td>temporal</td>
<td>positive</td>
<td>negative</td>
<td>binding</td>
<td>authors</td>
<td>comments</td>
</tr>
<tr>
<td>------------</td>
<td>------</td>
<td>-----------</td>
<td>------------------</td>
<td>----------</td>
<td>----------</td>
<td>----------</td>
<td>---------</td>
<td>---------</td>
<td>----------</td>
</tr>
<tr>
<td>19 kDa zein</td>
<td>pMS1</td>
<td>maize</td>
<td>nd</td>
<td>nd</td>
<td>-353/-312 (-300 element)</td>
<td>-353/-312 (-300 element)</td>
<td>Quayle and Feix 1992</td>
<td>Mutation of 2 bp in the core of the -300 element reduced activity.</td>
<td></td>
</tr>
<tr>
<td>19 kDa zein</td>
<td>pMS2</td>
<td>maize</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Langridge 1985</td>
<td>Transcription of zein in algal cells.</td>
<td></td>
</tr>
<tr>
<td>19 kDa zein</td>
<td>ZE19</td>
<td>maize</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Coniagio 1986</td>
<td>Transcription of zein in yeast cells.</td>
<td></td>
</tr>
<tr>
<td>19 kDa zein</td>
<td>ZE19</td>
<td>maize</td>
<td>-1415 to -430 (P1 promoter) (-430 to -7 (P2 promoter))</td>
<td>-1415 to -430</td>
<td>-1415 to -430</td>
<td></td>
<td>Quattrocchio 1990</td>
<td>ZE19 also expressed in anthers of maize and transformed Petunia.</td>
<td></td>
</tr>
<tr>
<td>19 kDa zein</td>
<td>ZE19</td>
<td>maize</td>
<td>-1470 to -426 (P1 -300 \text{ box}) (-425 to -301 (P2 -300 \text{ box}) (-300 to -7 (7-11-7 \text{ motif})) (-300 to -7 (5-11-5 \text{ motif}))</td>
<td>-1470 to -426 (P1 -300 \text{ box}) (-425 to -301 (P2 -300 \text{ box}) (-300 to -7 (7-11-7 \text{ motif})) (-300 to -7 (5-11-5 \text{ motif}))</td>
<td></td>
<td>Giovinazzo 1992</td>
<td>Activation by the DNA regions varied between protoplasts from different sources.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>19 kDa zein</td>
<td>Z19ab1</td>
<td>maize</td>
<td>nd</td>
<td>nd</td>
<td>-800 to +1*</td>
<td></td>
<td>Goldsbrugh 1986</td>
<td></td>
<td></td>
</tr>
<tr>
<td>19 kDa zein</td>
<td>Z19ab1</td>
<td>maize</td>
<td>nd</td>
<td>nd</td>
<td>-526 to +1</td>
<td></td>
<td>Boston 1987</td>
<td></td>
<td></td>
</tr>
<tr>
<td>19 kDa zein</td>
<td>Z19ab1</td>
<td>maize</td>
<td>nd</td>
<td>-850 to +1 (\text{not specific; low expression in leaves and seeds})</td>
<td>-850 to +1 (\text{not specific; low expression in leaves and seeds})</td>
<td></td>
<td>Ueng 1988</td>
<td>No zein protein detected, only low level of zein transcript.</td>
<td></td>
</tr>
<tr>
<td>19 kDa zein</td>
<td>Z19ab1</td>
<td>maize</td>
<td>nd</td>
<td>nd</td>
<td>-337 to -126 (-125 to +61)</td>
<td></td>
<td>Rousell 1988</td>
<td>Region -337/-125 contains-300 box and AT-rich repeats and is equivalent to -394 to -189 in Thompson et al 1990, below).</td>
<td></td>
</tr>
<tr>
<td>19 kDa zein</td>
<td>Z19ab1</td>
<td>maize</td>
<td>nd</td>
<td>nd</td>
<td>-483^ to -348^ (\text{context dependant}) (-347^ to -226^ (-300 \text{ box, AT-rich}) (-189^ to -114^ (\text{GAAT}) (-114 to -8 (\text{TATA}))</td>
<td></td>
<td>Thompson 1990</td>
<td>-300 box was mutated within -347 to -226 without affecting expression. ^ Numbering relative to translation start codon.</td>
<td></td>
</tr>
<tr>
<td>PROLAMINES</td>
<td>gene</td>
<td>organisms</td>
<td>seed specificity*</td>
<td>temporal</td>
<td>positive</td>
<td>negative</td>
<td>binding</td>
<td>authors</td>
<td>comments</td>
</tr>
<tr>
<td>------------</td>
<td>------</td>
<td>-----------</td>
<td>-------------------</td>
<td>----------</td>
<td>----------</td>
<td>----------</td>
<td>---------</td>
<td>---------</td>
<td>----------</td>
</tr>
<tr>
<td>21 kDa zein</td>
<td>pML1</td>
<td>maize</td>
<td>nd</td>
<td>nd</td>
<td>-2100 to +1*</td>
<td></td>
<td></td>
<td>Langridge 1983</td>
<td>Transcription of zein in yeast and HeLa cells. Gene is transcribed from two widely separated promoters regions</td>
</tr>
<tr>
<td>21 kDa zein</td>
<td>pML1</td>
<td>maize</td>
<td>nd</td>
<td>nd</td>
<td>-1240 to -1000</td>
<td></td>
<td></td>
<td>Brown 1986</td>
<td></td>
</tr>
<tr>
<td>22 kDa zein</td>
<td>22Z-2</td>
<td>maize</td>
<td>nd</td>
<td>nd</td>
<td>-307 to -284 (palindrome and RY, core site is ACGT)</td>
<td></td>
<td></td>
<td>Schmidt 1992</td>
<td>RY was within O2 footprint, but was not required for binding. G within ACGT was essential for binding.</td>
</tr>
<tr>
<td>22 kDa zein</td>
<td>22Z-2</td>
<td>maize</td>
<td>nd</td>
<td>nd</td>
<td>-295 to -286 O2 binding site is TCCACGTAGA</td>
<td></td>
<td></td>
<td>Ueda 1992</td>
<td>Base in bold were essential for binding and transactivation by O2</td>
</tr>
<tr>
<td>23 kDa zein</td>
<td>Z4</td>
<td>maize</td>
<td>nd</td>
<td>nd</td>
<td>-933 to +1^</td>
<td>-933 to +1^</td>
<td></td>
<td>Matzke 1984</td>
<td></td>
</tr>
<tr>
<td>23 kDa zein</td>
<td>Z4</td>
<td>maize</td>
<td>-174 to +1</td>
<td>-174 to +1</td>
<td>-886 to -174</td>
<td>-174 to -79</td>
<td></td>
<td>Schemthaner 1988</td>
<td></td>
</tr>
<tr>
<td>C hordein</td>
<td>Hor1-14</td>
<td>barley</td>
<td>-428 to +1</td>
<td>-431 to +1</td>
<td>nd</td>
<td>nd</td>
<td>Matzke 1990</td>
<td>Deletion of -300 core (15 bp) had no effect upon temporal/seed-specific or quantitative regulation.</td>
<td></td>
</tr>
<tr>
<td>B1 hordein</td>
<td>pBHR184</td>
<td>barley</td>
<td>-512 to +37</td>
<td>-512 to +37</td>
<td>nd</td>
<td></td>
<td>Matzke 1990</td>
<td>Little homology between S regions except for -300 element (30 bp), cereal CACA box and TATA box.</td>
<td></td>
</tr>
<tr>
<td>gliadin</td>
<td>gli-1</td>
<td>wheat</td>
<td>nd</td>
<td>nd</td>
<td>-592 to -440 (Z-DNA)</td>
<td>-218 to -142 (CACA)</td>
<td>-142 to -77 (CAAT + TATA)</td>
<td>Aryan 1991</td>
<td>-300 box was deleted without affecting level of expression. CACA element was located within a quantitative element.</td>
</tr>
<tr>
<td>low molecular weight glutenin</td>
<td>L.MWG-1D1</td>
<td>wheat</td>
<td>-326 to +30</td>
<td>-326 to +30</td>
<td>nd</td>
<td>nd</td>
<td>Colot 1987</td>
<td>Positive element contains two-300 boxes.</td>
<td></td>
</tr>
<tr>
<td>high molecular weight glutenin</td>
<td>Glu-D1-2</td>
<td>wheat</td>
<td>-326 to +30</td>
<td>-326 to +30</td>
<td>nd</td>
<td>nd</td>
<td>Colot 1989</td>
<td>Positive element contains two-300 boxes.</td>
<td></td>
</tr>
<tr>
<td>PROLAMINES</td>
<td>gene</td>
<td>organisms</td>
<td>seed specificity*</td>
<td>temporal</td>
<td>positive</td>
<td>negative</td>
<td>binding</td>
<td>authors</td>
<td>comments</td>
</tr>
<tr>
<td>----------------------------</td>
<td>------------</td>
<td>-----------</td>
<td>-------------------</td>
<td>-------------------</td>
<td>--------------</td>
<td>--------------</td>
<td>---------</td>
<td>-------------</td>
<td>--------------------------------------------------------------------------</td>
</tr>
<tr>
<td>high molecular weight glutenin</td>
<td>Glu-D1-2</td>
<td>wheat</td>
<td>-375 to -148 (-163 to -148 endosperm specificity)</td>
<td>-375 to -168</td>
<td>-234 to-196 (minor)</td>
<td>-186 to -163</td>
<td>-163 to -148</td>
<td>Thomas 1990</td>
<td>Enhancer region -186 to -148 has repeats of TTGCT and GCTCC, but no -300 box.</td>
</tr>
<tr>
<td>high molecular weight glutenin</td>
<td>Glu-D1-2</td>
<td>wheat</td>
<td>-433 to -5</td>
<td>-433 to -5</td>
<td></td>
<td></td>
<td></td>
<td>Robert 1989</td>
<td></td>
</tr>
<tr>
<td>high molecular weight glutenin</td>
<td>Glu-1Dx</td>
<td>wheat</td>
<td>-277 to +39 (10bp changes from Glu-1Dx to Glu-1Ay promoter silences expression)</td>
<td>-277 to +39</td>
<td></td>
<td></td>
<td></td>
<td>Hallorde 1989</td>
<td>Consensus of the -300 box, was not mutated by any of the bp changes.</td>
</tr>
<tr>
<td>glutelin</td>
<td>Gt 1</td>
<td>rice</td>
<td>-154 to +1</td>
<td>-154 to +1</td>
<td></td>
<td></td>
<td></td>
<td>Mural 1991</td>
<td></td>
</tr>
<tr>
<td>glutelin</td>
<td>Gt 3</td>
<td>rice</td>
<td>-980 to +96, but also high expression in leaves</td>
<td>-980 to +96</td>
<td></td>
<td></td>
<td></td>
<td>Leisy 1989</td>
<td></td>
</tr>
<tr>
<td>glutelin type II</td>
<td>pREE771</td>
<td>rice</td>
<td>-441 to +1*</td>
<td>-441 to +1*</td>
<td>-441 to -237</td>
<td></td>
<td></td>
<td>Takaiwa 1991</td>
<td>&quot;AACA&quot; glutenin motif between -78/-65 (AAACAAACTCTATC). Four repeats between -540 and -192, of a sequence similar to -300 core.</td>
</tr>
<tr>
<td>glutelin type II</td>
<td>pREE771</td>
<td>rice</td>
<td></td>
<td></td>
<td></td>
<td>-120 to -130 CTTTCGTTGAC</td>
<td></td>
<td>Takaiwa 1990</td>
<td>Binding protein was not present in leaf or root nuclear extracts. Binding activity of the binding protein correlated with the mRNA levels of the gene during seed maturation.</td>
</tr>
</tbody>
</table>
Table 1.2 Frequency of putative seed-specific motifs in plant promoters.

A. Frequency of putative seed-specific motifs in genes that are highly expressed in seeds, including seed storage protein genes and related genes (globulins, prolamines, 2S albumins and lectins) and other genes that are highly expressed in seeds, during late embryogenesis and/or early germination (other).

B. Frequency of putative seed-specific motifs in genes that are not highly expressed in seed. The gene types have been separated into those encoding for vegetative storage, histone, heatshock and other proteins. The expected frequencies are calculated for the probability of each motif within 300 bp of random sequence.

The 5' sequences that were used all contained 100 bp or more and came from a database by compiled by Leon Dure and other published sequences. The results for individual sequences are listed in Appendix I. Observed frequencies that were significantly higher than the expected frequencies (\( \chi^2 \)-test with Yates correction, at 5%) are in bold.

- Octanucleotide box: GCCACc/tTC
- Vicilin box I: GCCACnTCAATTTt/-GTnCATTTnAACACnCGTC
- Vicilin box II: nTTCATnAAA TTCAAACAAAA
- Legumin box: TCCATAa/gCCATG CAa/tGa/cTGAAGAATGTC
- -300 box: TGTAAAG

Significant at the 10% level.

Four sequences were excluded because they contained less than 300 bp, and so were unlikely to contain the -300 box.

Patatin genes
### A. Seed-Expressed Genes

<table>
<thead>
<tr>
<th>Motif Frequency (%)</th>
<th>Expected</th>
<th>Observed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Legumes</td>
<td>Other</td>
<td>Cereals</td>
</tr>
<tr>
<td>globulins</td>
<td>albumins</td>
<td>other</td>
</tr>
<tr>
<td>7S</td>
<td>11S</td>
<td>2S</td>
</tr>
<tr>
<td>lectins</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### Motifs and Mismatches

<table>
<thead>
<tr>
<th>Motifs</th>
<th>Mismatches</th>
<th>7</th>
<th>11</th>
<th>1</th>
<th>7</th>
<th>3</th>
<th>6</th>
<th>6</th>
<th>2</th>
<th>18</th>
<th>11</th>
</tr>
</thead>
<tbody>
<tr>
<td>CATGCATG</td>
<td>1</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>86</td>
<td>33</td>
<td>33</td>
<td>0</td>
<td>50</td>
<td>39</td>
<td>45</td>
</tr>
<tr>
<td>CACA</td>
<td>0</td>
<td>86</td>
<td>45</td>
<td>0</td>
<td>71</td>
<td>67</td>
<td>17</td>
<td>83</td>
<td>50</td>
<td>67</td>
<td>45</td>
</tr>
<tr>
<td>ANCCCA</td>
<td>0</td>
<td>71</td>
<td>45</td>
<td>0</td>
<td>100</td>
<td>17</td>
<td>66</td>
<td>50</td>
<td>44</td>
<td>90</td>
<td>90</td>
</tr>
<tr>
<td>octanucleotide box a</td>
<td>0</td>
<td>86</td>
<td>73</td>
<td>0</td>
<td>43</td>
<td>0</td>
<td>17</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>vicilin box I b</td>
<td>3</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>vicilin box II c</td>
<td>2</td>
<td>71</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>legumin box d</td>
<td>2</td>
<td>0</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>-300 box e</td>
<td>0</td>
<td>0</td>
<td>27</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>17</td>
<td>33</td>
<td>0</td>
<td>73</td>
<td>0</td>
</tr>
</tbody>
</table>
### B. Non-Seed Genes

<table>
<thead>
<tr>
<th>Motif Frequency (%)</th>
<th>Expected</th>
<th>Observed</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Storage</td>
<td>Histone</td>
</tr>
<tr>
<td></td>
<td>dicot</td>
<td>dicot</td>
</tr>
<tr>
<td>Motifs</td>
<td>Mismatches</td>
<td>2</td>
</tr>
<tr>
<td>CATGCATG</td>
<td>1</td>
<td>11</td>
</tr>
<tr>
<td>CACA</td>
<td>0</td>
<td>15</td>
</tr>
<tr>
<td>ANCCCA</td>
<td>0</td>
<td>59</td>
</tr>
<tr>
<td>octanucleotide box</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>vicilin box I</td>
<td>3</td>
<td>&lt;1</td>
</tr>
<tr>
<td>vicilin box II</td>
<td>2</td>
<td>&lt;1</td>
</tr>
<tr>
<td>legumin box</td>
<td>2</td>
<td>&lt;1</td>
</tr>
<tr>
<td>-300 box</td>
<td>0</td>
<td>4</td>
</tr>
</tbody>
</table>
CHAPTER 2. DELETION ANALYSIS OF THE 5' REGION OF PSA-V.

2.1 Introduction

Gene transfer studies with *Psa-v*, a vicilin gene from pea, indicated that 2500 bp of the 5' region was sufficient for seed-specific and temporally-correct expression in transgenic tobacco (Higgins *et al.*, 1988). Comparison of the 5' region of vicilin and other seed storage protein genes have identified a number of conserved DNA motifs, that might contain elements that regulate gene expression. The upstream region of *Psa-v* contains many of the conserved sequences that have been observed in seed-specific genes from legumes (Figure 2.1). Two of these sequences, an octanucleotide motif (Rerie *et al.*, 1992) and a CACA element (Goldberg, 1986) are contained within vicilin box I, located approximately 130 bp upstream from the transcription start site of vicilin genes from legumes (Gatehouse *et al.*, 1986; Bown *et al.*, 1988; Newbigin, 1988; Higgins *et al.*, 1988; Rerie *et al.*, 1992; Figure 1.3). The 5' region of *Psa-v* also contains a CATGCATG motif at -96 (also called an RY repeat, Dickinson *et al.*, 1988) and a second vicilin-specific element, termed vicilin box II (Newbigin, 1988) located approximately 150 bp from the transcription start site. Previous data, however, suggest that these conserved sequences are not sufficient for expression and that additional upstream DNA sequences are required. Newbigin (1988; Appendix II), made chimeric gene constructs containing either -2500/-49 or -228/-49 of *Psa-v*, ligated to the TATA box region of the CaMV 35S promoter plus the coding region of the gene encoding chloramphenicol acetyl transferase (CAT) and the 3' untranslated region of the nopaline synthase gene.)
Although both these Psa-v fragments contain the CATGCATG element, vicilin boxes I and II conserved sequences, no expression could be detected in transgenic tobacco when the region -228/-49 was used to promote CAT gene expression. The gene construct containing -2500/-49 of Psa-v, however, conferred high-level gene expression that was both temporally-regulated (Newbigin, 1988) and seed-specific (Appendix II). This indicated that there is at least one important regulatory element between -2500 and -229 of Psa-v.

Functional analyses of other seed storage protein gene promoters (Shirsat et al., 1989; Baumlein et al., 1991; Rerie et al., 1991) also suggest that the conserved sequences may not be sufficient for high-level expression. In contrast, however, analysis of the gene CGy1, which encodes α',β-conglycinin in soybean, found that sequences downstream of -159, which includes vicilin box I and CATGCATG elements, conferred seed-specific expression in transgenic petunia (Chen et al., 1986). This expression was low, however. Full expression required the presence of the region -257/-160. Thus, these results support the general hypothesis that other, upstream, DNA sequences (perhaps in addition to the conserved sequences) are required for high level expression of seed-specific genes.

This chapter describes the results of experiments that were designed to locate cis-acting elements within the 5'-flanking sequences of Psa-v. A set of Psa-v gene constructs with progressive deletions of the 5' region were made and transferred to tobacco. The seeds and vegetative organs of the transformed plants were assayed for the presence of vicilin by immuno-blotting (Western blotting). These constructs were designed to:

- define the minimal amount of 5' sequence that was required for detectable expression of Psa-v
- determine whether this minimal sequence contains elements sufficient for control of seed-specific expression and/or temporal regulation
- establish whether additional (positive) elements exist upstream of the minimal promoter.
Figure 2.1
Conserved and repeated sequences in the 5' sequence of *Psa-v*.

Numbering is relative to the transcription start site (Higgins *et al.*, 1988), with a six base pair correction at position -296, in lower case. Numbers indicating 5' deletion end-points are shown in bold. The AT-rich repeats between -550/-458, -462/-370 are underlined, with double underlining at the overlap. The non-matching bases in the two repeats are shown in lower case. A small interrupted palindrome, ↔, centred at -285, is overlined, with the inverted repeats in italics. Vicilin box II (Newbigin, 1988) is highlighted. The 33 bp vicilin box I (Rerie *et al.*, 1992) is double underlined and the extended box I (Newbigin, 1988) is underlined with a single line. The conserved octanucleotide ▼, CACA▼, CATGCATG ▲, CAAT and TATA boxes are shown in bold.

Figures 2.1 & 2.2 should be exchanged.
2.2. Materials and Methods

2.2.1. Construction of 5' deletion series of Psa-v

A series of deletions was made by digesting the 5' region of the Psa-v coding region with Nar I and Eco RI, and the fragments were ligated with pAGA492 (Fig. 2.2A). The deleted sequences were then sequenced using a 3' reverse primer and a 5' forward primer. The results showed that the deletion series was successful.

Figure 2.2
**2.2 Materials and Methods**

### 2.2.1 Construction of 5' deletion series of *Psa-v*.

A series of *Psa-v* gene constructs was made in which the 5' region was progressively deleted. Plasmid pEN2, which contains the vicilin gene, *Psa-v*, with 2.5 kb of the 5' region, 2.2 kb of coding region and .3 kb of 3' region had been used to construct the gene construct pAGB1 (Higgins *et al.*, 1988). Gene construct pAGB1, with the full-length *Psa-v* gene in a binary vector, is referred to as \( \Delta-2500 \) in this thesis.

Plasmid pCW66 was made by ligating a 5.5 kb *Eco RI* fragment (containing all of *Psa-v*) from pEN2 into pUC118, and then extra restriction sites were added by *in vitro* mutagenesis to make plasmid pCW66 (Wandelt *et al.*, 1992). These new restriction sites, *Cla I* (C), *Nco I* (Nco), *Bam HI* (B) and *Xho I* (Xho) were located at sites corresponding to the start sites for transcription, translation, peptide cleavage site and 3' end respectively and are underlined in Figure 2.2A.

An initial set of 5' deletion constructs was made by isolating the 5' region fragments *Hinc II-Nco I*, *Bcl I-Nco I*, *Ssp I-Nco I*, *Acc I-Nco I*, *Dde I-Nco I*, *Nde I-Nco I*, from pCW66. These sites, on the 5' side, correspond, respectively, with positions 1470, 1038, 767, 418, 300 and 96 bp upstream of the transcription start site of the published sequence Higgins *et al.*, 1988). Further sequencing of this region indicated the presence of 6 additional base pairs at position -296, and so these deletions were actually at positions -1476, -1042, -773, -424, -306, and -96. These 5' fragments were blunted at the 5' end and ligated into the pCW66 *Sma I-Nco I* fragment (containing all of the plasmid except the *Psa-v* 5' region). This generated plasmids pCW81, 78, 70, 74, 71 and 81, respectively.

Each of these 5' deletion constructs were then digested with *Nar I* and *Eco RI*, and the fragments were ligated into the *Cla I-Eco RI* fragment of the binary vector pGA492 (An, 1987a), generating plasmids pCW84, 80, 75, 79, 76 and 83 (see Figure 2.2A). These are termed \( \Delta-1476, \Delta-1042, \Delta-773, \Delta-424, \Delta-306 \) and \( \Delta-96 \), respectively. Figure 2.2B shows the location of the conserved and repeated sequences within the 5' regions of each of these gene constructs.
2.2 A. Structure of *Psa-v* deletion constructs.

The construction of pEN2 and the binary vector pAGAB1, containing the full length *Psa-v* gene is described in Higgins *et al.*, (1988). Plasmid pCW66 was made by the ligation of a 5.5 kb *Eco RI* fragment from pEN2, containing the *Psa-v* gene, into the *Eco RI* site of pUC118, followed by *in vitro* mutagenesis, to add extra restriction sites at the start sites for transcription, translation, the transit peptide cleavage site and the 3' region (Wandelt *et al.*, 1992).

Restriction sites in pCW66 are indicated, from left to right: *Nar I* (Nar), *Hind III* (III), *Hinc II* (some of which are also *Hpa I* sites, [H]), *Xba I* (X), *Bam HI* (B), *Ssp I* (S), *Eco RI* (E), *Bcl I* (Bcl), *Acc I* (A), *Dde I* (D), *Nco I* (N), *Cla I* (C), *Xho I* (Xho). Sites that were added by *in vitro* mutagenesis are underlined.

A series of 5' deletion mutants were made from pCW66, as described in chapter 2.2.1. The location of the restriction sites in the 5' region that were used to generate the preliminary (pUC118) plasmids, pCW81, 78, 70, 74, 71 and 82, at positions -1476, -1044, -773, -424, -306 and -96 of the *Psa-v* gene, respectively, are indicated. The *Nar I* - *Eco RI* fragments of each of these plasmids were ligated into the *Cla I* and *Eco RI* sites of binary vector, pGA492, generating plasmids pCW84, 80, 75, 79, 76, and 83, as indicated. The length of the *Psa-v* gene in plasmid pAGAB1, (which otherwise differs from that in the CW plasmids by only the underlined restriction sites), is also indicated.

The location of the left (B_L) and right (B_R) borders of the T-DNA and the Kanamycin resistance (*kan*), promoterless Chloramphenicol Acetyl Transferase (*cam*) and Tetracycline resistance (*tet*) genes are indicated as Kan R, CAT and Tet R, respectively. The kanamycin resistance gene was modified for expression in plants by addition of the 5' and 3' region of the nopaline synthase gene to the *kan* coding region.

Stippled boxes represent *Hpa I* - *Xba I* fragments of the coding region that were used as probes for Southern blots (this chapter) and Northern blots (Chapter 4).
-773 ATT AGTTTTAAAT GTTTTTATTA

-750 AAAATAAATT GTATTTTTTA TCCTTATTAC TCCATCTTTCT CTTTACTAAT

-700 AATGCCCTGA GAATACGGAT ATATTTGGAC ATGACAAATA TAACCGTTTTT

-650 AATAAATTTAT CATCTTTTG TGATGAGAATT ATGTAACCTAT ATTTTTTTTTA

-600 ATTTGAGATT GTAATAAAAG TAGATTATAA TATATATTAT ATTTTTTTTTT

-550 TTTATAATATA TAAAATTAAG TATAATATAT gTaAGATAA CCgATAAAATA

-4621-4581

-500 ATAGATAAT AATAAAATGA CATAATATA cAACTATTT TTTATATATT

-450 AAgTAGATAA ATAGATAAT GAAGTAGACG GATAAAATAT AGATAGATAA

-400 TTTAAATGACg TATATgTAGA ATTACATTTT TCACAGACg AGTACAACCC

-350 TATGCACTTC TAAGTGCAAG TTTATGGAGT TAGTCCCATG TCTTAGAGct

-300 ggatCTTGTG TTTAGGATA CAACACTTGT TAAAATTCgT TAGTCaATTC

-250 ATTAATTcat ATACACATGG CGAAGACAA TAAATAAGCA TCCTCTTTTT

-VICILIN BOX II-

-200 CCATAAGAT GTCCCAAATTC ATCAAAATTCA AACAACACTC CACCACCCCA

-oct ▼

-150 GTATAAGTCT CTTCATTTTG CCACTTGCAAT TTTTGCTATT TTAACACCCG

-961 CATGCATG ▲

-100 TCCATATGCA TTGCAACACA TGGCCAACCTG TGCTCGCATG TTAATTATAT

TATA

- 50 AGTTTTTATT TTTATATCTA TAAAATACCT CATCTCAGTC TACTTATTTC

Figure 21.
Figure 2.2
*Psa-v 5' deletion gene constructs.*

2.2 B. Location of conserved motifs within the 5' deletion constructs.

Deletion end-points are relative to the start site of transcription, and are all to scale, except the -2500 deletion. The location of conserved and repeated DNA sequences within the 5' region of *Psa-v* are shown; from right to left: TATA box (T), CAAT box (C), CATGCATG element (▲, mismatched nucleotide in lower case), vicilin box I (overlined, containing a CACA box, ▼ and an octanucleotide box, ▼), vicilin box II (▼,), a palindrome (↔) and AT-rich repeats (►). Positions are relative to the start site of transcription.
2.2.2 Transformation of plants

The binary plasmids containing 5' deletions of Psav were transferred from E. coli (HB101) into A. tumefaciens (strain AG1111) and LBA4404 (Hoeschele et al., 1990) by transformation method using Agrobacterium, 2.5 pl. Bacto-yeast extract, 2.5 pl. water, and 2 pl. 70% ethanol. Binary plasmids were transformed into the plants with the same Horsfall technique. The plasmids were first incubated in Piggirea 0.1 pl. and then incubated for 2 to 3 minutes and were transformed in hormone-free 1/10 MS medium (2 g/L sucrose, 100 mM MES, 0.1 mg/L 2,4-D, 0.1 mg/L NAA, 125 mM KNO₃, 0.22 mM MgSO₄·7H₂O, 0.005 mM KCl, 0.01 mM H₃BO₃, 0.022 mM MnCl₂·4H₂O, 0.022 mM ZnSO₄·7H₂O, 0.005 mM H₃BO₃, 0.01 mM HCl). The plates were incubated in the dark at 20°C in the dark for 4 weeks. After that, some plasmids containing shoot induction medium (40 mg/L KNO₃, 60 mg/L succinic acid at pH 6.5, 1 mg/L benzylaminopurine at 1 mg/L) and 1 mg/L of the antibiotic were used to kill the Agrobacterium and kanamycin (100 mg/L) to select the transformed plant cells. Plants that formed roots were transferred to soil and grown in the greenhouse. Plants were tagged when the petals opened, which correlates with anthesis, and the flowers were collected at different developmental stages.
2.2.2 Transformation of plants.

The binary plasmids containing 5' deletions of Psa-v were transferred from *Escherichia coli* (HB101) into *A. tumefaciens* (strains AGL1 [Lazo *et al.*, 1991] and LBA4404 [Hoekema *et al.*, 1983]) by triparental mating, using the helper plasmid RK2013, as described by Ditta *et al.* (1980). Transconjugants containing the binary vector were selected on LBMG medium (20 g/l, agar; 5 g/l, Bactotryptone; 2.5 g/l, Bacto yeast extract; 87 mM, NaCl; 27.5 mM, mannitol; 7.82 mM, glutamic acid; KH$_2$PO$_4$, 1.84 mM; 0.41 mM, pH 7.0 MgSO$_4$.7H$_2$O) containing rifampicin (50 mg/l) and tetracycline (2 mg/l). The *A. tumefaciens* transconjugants were used to transform tobacco leaf tissue, and plants were regenerated from transformed cells, essentially as described in Higgins *et al.* (1988). In brief, leaf pieces were soaked in the *Agrobacterium* solution for 2 to 5 minutes and were then transferred to hormone free MS medium (revised basal medium, Murashige and Skoog, 1962): inorganic salts [mM]: 41.2 NH$_4$NO$_3$; 18.8 KNO$_3$; 3.0 CaCl$_2$.2H$_2$O; 1.5 MgSO$_4$.7H$_2$O; 1.25 KH$_2$PO$_4$; 0.2 Na$_2$-EDTA; 0.1 FeSO$_4$.7H$_2$O; 0.1 H$_3$BO$_3$; 0.1 MnSO$_4$.4H$_2$O; 0.03 ZnSO$_4$.7H$_2$O; 0.005 KI; 0.001 Na$_2$MoO$_4$.2H$_2$O; 0.0001 CuSO$_4$.5H$_2$O; 0.0001 CoCl$_2$.6H$_2$O and organic substances [mg/l] 30 000 sucrose; 10 000 agar; 100 myo-Inositol; 2.0 glycine; 0.5 nicotinic acid; 0.5 pyridoxine-HCl; 0.1 thiamine.HCl. The next day, the leaf pieces were transferred to MS9 plates containing shoot induction medium (MS medium with indole acetic acid at 0.5 mg/l, benzyl-aminopurine at 1 mg/l) and the antibiotics cefotaxime (300 μg/ml, to kill the *Agrobacterium*) and kanamycin (100 μg/m, to select the transformed plant cells). Shoots formed within 2 to 3 weeks; these were transferred to MS medium containing the same antibiotics, but no hormones, for rooting. Plants that formed roots were transferred to soil and grown in the glasshouse. Flowers were tagged when the petals opened, which coincides with anthesis, and seeds were collected at different developmental stages.
2.2.3 DNA isolation and Southern blotting of genomic DNA.

DNA was isolated from young tobacco leaves using a method modified from Shure et al. (1983). Approximately 2 g of leaf tissue was frozen with liquid nitrogen and then ground to a powder with a mortar and pestle. The powder was added to a 50 ml polypropylene Falcon tube containing 6 ml of urea extraction buffer (7 000 mM urea, 310 mM NaCl, 50 mM Tris-HCl, pH 8, 20 mM EDTA, pH 8.0, 1% (w/v) sarkosine). Five ml of phenol: chloroform: isoamylalcohol (25:24:1) was added and the mixture was shaken vigorously for 5 min and then centrifuged at 5,000 g for 10 min. Approximately 5 ml of the aqueous phase was transferred to a fresh tube. The nucleic acid was precipitated by the addition of 1 ml of 4400 mM ammonium acetate, pH 5.2 and 7 ml of isopropanol. This mixture was inverted twice, left at room temperature for 2 min and the DNA was spooled onto a blunt Pasteur pipette. The DNA was then dissolved in 500 µl of TE (10 mM Tris-HCl pH 8.0, 1 mM EDTA) in an Eppendorf tube and re-precipitated with 100 µl of 4400 mM ammonium acetate and 700 µl of isopropanol, respooled, and washed with 80% (v/v) ethanol. Finally, the DNA was dissolved in 400 µl of TE, overnight, at 4oC, and any undissolved material was removed by centrifugation before it was digested with restriction enzymes.

After digestion with Eco RI and Bam HI, the DNA was quantified by absorbance at 260 nm. Five µg were fractionated on 1.2 % (w/v) agarose gels and transferred to Hybond-N (Amersham) membrane. The gels were stained with ethidium bromide and photographed to confirm that each cell was evenly loaded. The gels were washed in 0.25 N HCl for 10 min to nick the DNA, denaturing solution (1.5 M NaCl, 0.5 M NaOH) for 30 min, and neutralised in 1.5 M NaCl, 0.5 M Tris pH 7.0 for 30 min. The DNA was transferred overnight, to nylon membrane (Hybond-N, Amersham) using capillary blotting; using a bottom reservoir of paper towelling saturated with 20 X SSC (1x SSC = 150 mM NaCl, 15 mM sodium citrate). The damp membrane was UV irradiated for 2 min with a germicidal 15 W tube at a distance of 12 cm to crosslink the DNA to the membrane (Khandjian, 1987).

The membranes were incubated at 42oC for 3 h in Khandjian's solution (0.1 ml per cm2 of membrane(Khandjian (1987): 50 mM Tris-HCl pH 7.5, 1 M NaCl, 50% (v/v) formamide, 10x Denhart's solution (Denhart, 1968); 0.2% (w/v) each of Ficoll, polyvinylpyrrolidone and bovine serum albumin [Pentax fraction V], in water ), 10% (w/v) dextran sulphate, 1% (w/v) SDS, 0.1% (w/v) sodium
pyrophosphate, 10 µg/ml salmon sperm DNA). A radiolabelled probe using a 305 bp \textit{Hpa I- Xba I} fragment containing a portion of the \textit{Psa-v} coding region (Figure 2.2a) was prepared by random-priming, using $^{32}$P-dCTP (Bresatec, Adelaide), to a specific activity between $10^8$ - $10^9$ cpm/µg DNA. The probe was then denatured by the addition of NaOH to 0.1 M and incubation at 37°C for 5 min, and added to the pre-incubated membrane in the solution.

Following overnight hybridisation at 42°C, the membranes were washed twice in 2x SSC with 0.1% (w/v) SDS, 0.1% (w/v) sodium pyrophosphate, at room temperature for 15 min, and then once in 0.1x SSC, 0.1% (w/v) SDS, 0.1% (w/v) sodium pyrophosphate, at 65°C for 30 min. The membrane was exposed to Kodak X-ray film for 4 days at -80°C with an intensifying screen.

### 2.2.4 Isolation and detection of vicilin in transgenic plants.

Vicilin was assayed in different organs of transgenic plants using western blots as described in Higgins and Spencer (1991). Approximately 1 g of leaf, stem or root tissue, or 100 to 250 mg of seed were ground in a mortar or Eppendorf tube in 3-4 volumes of grinding buffer (100 mM TES, 200 mM NaCl, 1.0 mM EDTA, 1 mM PMSF, 2% (v/v) β-mercaptoethanol). The mixture was centrifuged at 10,000 g for 15 min and the supernatant collected. The concentration of protein in the sample was determined using Bradford reagent (Bradford, 1976). Vicilin was extracted from 0.4 to 1.5 ml of the supernatant by incubation with 20 µl of Sepharose 4B beads which had been linked covalently to sheep immunoglobulin G specific for vicilin (Badenoch-Jones \textit{et al.}, 1981; Spencer \textit{et al.}, 1980). The extract and beads were tumbled gently for one hour and then the Sepharose beads were collected by centrifugation, washed 5 times in a solution containing 450 mM NaCl, 30 mM Na-phosphate (pH 7.4) and 1% (v/v) Tween-20 and resuspended in PAGE sample buffer (0.125 M Tris-HCl pH 6.7, 2% (v/v) SDS, 10% (v/v) glycerol, 0.003% (w/v) bromophenol blue and 5% (v/v) β-mercaptoethanol). The mixture was heated at 90°C for 10 min. and an aliquot equivalent to 0.5 mg of soluble protein in the original extract was fractionated on an 18% SDS-polyacrylamide gel (Laemmli, 1970; Spencer \textit{et al.}, 1980).

Protein was electro-blotted onto nitrocellulose membrane as described by Spencer \textit{et al.} (1983), using semi-dry blotting. The gel was soaked in electrobolt buffer (25 mM Tris, 192 mM glycine, 20% ethanol, 0.02% (w/v) SDS,
pH 8.3) for 20 min and then a blotting "sandwich" was made, with 6 layers of 3 mm Whatmann blotting paper soaked in anode solution 1 (300 mM Tris, 20% (v/v) methanol, pH10.4) on the bottom graphite slab, followed by 3 layers of paper soaked in anode solution 2 (25 mM Tris, 20% (v/v) methanol, pH 10.4), then wet nitrocellulose (soaked in water), the gel, 9 layers of paper soaked in cathode solution (40 mM 6-amin-n-hexanoic acid, 20% (v/v) methanol, pH 7.5) and finally the cathode (graphite) plate. A constant current of 0.8 mA/cm² was applied to the apparatus for 1 hour. The nitrocellulose membrane was then stained in amido black stain (0.01% (w/v) naphthalene blue black 12 B, 0.45% (v/v) methanol, 1% (v/v) acetic acid) for 5 min, to monitor the transfer of the proteins from the gel to the membrane.

Vicilin was detected colorimetrically by incubation with rabbit anti-vicilin serum, followed by goat anti-rabbit IgG (Promega, Madison WI), conjugated to alkaline phosphatase (Bio-Rad, Richmond, Cal., USA). Briefly, the membranes were blocked for 30 min in blocking solution (5% (w/v) Diploma skim milk powder in Tris buffered saline (TBS): 500 mM NaCl, 20 mM Tris-HCl, pH7.4), then washed 3 times in TBS with 0.05% Tween-20 (TTBS) for 5 min. They were incubated with the anti-vicilin antibody for one hour (crude serum was diluted 1/200 in TTBS plus 1% (w/v) skim milk powder before use), washed 3 times in TTBS for 5 min. Next, the membranes were incubated with the anti-rabbit antibody (diluted 1/6666 in TTBS plus 1% (w/v) skim milk powder and washed 3 times for 5 min in TTBS. The colorimetric reaction was developed by incubation of the membranes in bromo-chloro-indolyl-phosphate at 0.15 mg/ml, nitroblue tetrazolium at 0.30 mg/ml, 100 mM NaHCO₃, 1 mM MgCl₂, pH 9.8, for 1 to 10 min.

The concentration of vicilin within the transgenic seeds was estimated by densitometry, using a purified pea vicilin as a standard. Negatives were scanned using a Joyce-Loebel double beam micro-densitometer (MK111 CS). The relative contributions of individual bands were monitored by excising and weighing the peaks.
2.3 Results

2.3.1 Analysis of transformed of tobacco plants

Southern blot hybridisation was used to determine whether the regenerated transgenic plants contained the Psa-v gene constructs. For each construct, DNA from six to nine plants was analysed to assess the presence and integrity of the transgenes. The results from a representative blot are shown in Figure 2.3. This DNA was digested with the enzymes Eco RI and Bam HI, which, if the constructs were present within the transgenic host, would all produce a 2.7 kb fragment (containing the coding and 3' regions), with the exception of the Δ-2500 construct, which should release a 5.5 kb fragment (containing the entire gene).

As shown in Figure 2.1 (first lane), a plant transformed with the Δ-2500 construct contained a 5.5 kb fragment that hybridised with the probe, indicating that this plant contained at least one copy of the construct. Three of the four plants putatively transformed with construct Δ-424 (lanes 2-5) and all of those transformed with construct Δ-1476 (lanes 7-11) contained hybridising bands of 2.7 kb, indicating that the Psa-v constructs were present in these plants.

No consistent correlation between copy number and gene expression was observed (data not shown). Since the range of copy number between constructs was quite uniform, the plants were assessed for gene expression without normalising for gene copy number. (To avoid unknown gene copy effects in non-seed tissues, however, the plants that were analysed for expression of vicilin in leaves, stems and roots each contained only 1-2 copies of their respective constructs).

Some plants, such as Δ-424-D (third lane), did not contain any hybridising bands. These plants also failed to re-root on media containing kanamycin, indicating that they were untransformed "escapes". Lane c, which contained tobacco DNA from a plant transformed with the binary vector pGA492, did not, as expected, show any bands that hybridised with the Psa-v probe.
Figure 2.3
Southern blot analysis of genomic DNA from tobacco plants transformed with 5' deletion constructs.

Autoradiograph of a representative genomic Southern blot. Each lane contained 5 µg of DNA isolated from tobacco transformed with 5' deletion constructs -2500 (lane 1), -424 (lanes 2 to 5), pGA492 (C, lane 6) and -1476 (lanes 7 to 11). The DNA was digested with the enzymes Eco RI and Bam HI. The expected radioactive fragments were 5.5 kb for the -2500 construct and 2.7 kb for the -424 and -1476 constructs.
5' deletion constructs

<table>
<thead>
<tr>
<th>Plant</th>
<th>Size (kb)</th>
<th>△-2500</th>
<th>△-424</th>
<th>C</th>
<th>△-1476</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>5.5</td>
<td><img src="image" alt="Image" /></td>
<td><img src="image" alt="Image" /></td>
<td><img src="image" alt="Image" /></td>
<td><img src="image" alt="Image" /></td>
<td><img src="image" alt="Image" /></td>
</tr>
<tr>
<td>2.7</td>
<td><img src="image" alt="Image" /></td>
<td><img src="image" alt="Image" /></td>
<td><img src="image" alt="Image" /></td>
<td><img src="image" alt="Image" /></td>
<td><img src="image" alt="Image" /></td>
</tr>
</tbody>
</table>

Western blots of protein extracted from the mature seed of each transgenic plant were used to assess the level of expression of violin obtained from each of the different 5' deletion constructs. In Figure 2A-D, each lane contained protein from a pooled sample of seeds from the mature seed of 6-9 plants transformed with a particular gene construct. Individual analysis of the seed from each plant detected violin in the seed of all plants which had been transformed with gene constructs containing 506 bp or 806 bp of the 5' region of violin. None of the seed from plants containing constructs lacking 2500 bp or 424 bp of the 5' region of violin contained detectable levels of violin protein. Violin was not detected in any of the seeds of plants containing constructs lacking -1476 bp of the 5' region of violin. Thus, the region downstream of position -96 contains a regulatory element that controls seed-specific expression of the violin gene. However, since the removal of sequences between -306 and -96 did not affect tissue specificity, it was not possible to infer whether the seed-specific regulatory element(s) are within the region -306 to -96 or are downstream of -96.

The locations of conserved and repeated sequences within the 5' region of the violin gene constructs are shown in Figures 2.1 and 2.2b. The seed expression region between positions -96 and -97, which has at least one important regulatory element, contains violin box II and half of violin box I, including the octanucleotide and CAAC motif. This region also includes a short palindromic sequence. The region downstream of position -96 contains the CATGCATG, CAAT, and TATA elements.
2.3.2 Seed-specific expression

Western blots of protein extracted from the mature seed of each transformed plant were used to assess the level of expression of vicilin obtained from each of the different 5' deletion constructs. In Figure 2.4D, each lane contained protein from a pooled samples of extracts from the mature seeds of 6-9 plants transformed with a particular gene construct. Individual analysis of the seed from each plant detected vicilin in the seed of all plants which had been transformed with gene constructs containing 306 bp or more of the 5' region of Psa-v (lanes -2500, -1476, -1042, -773, -424, -306). None of the seed from plants containing construct Δ-96, however, contained detectable levels of vicilin (Figure 2.4D lane -96 and data not shown). Thus, the region downstream of -306 contains sufficient cis-acting elements for detectable expression in seeds, and at least one element (termed the seed expression region) that regulates the level of expression is located between positions -306 and -96.

To determine whether the regulatory elements downstream of -306 were sufficient for controlling the seed-specific expression of Psa-v, vegetative organs of representative plants transformed with each gene construct were analysed for the presence of vicilin (Figure 2.4A-C). Vicilin was not observed in the leaves, stems or roots of any of these plants. Thus, the region downstream of position -306 does indeed contain regulatory elements that control seed-specific expression. However, since the removal of sequences between -306 and -96 did not affect tissue specificity, it was not possible to infer whether the seed-specific regulatory element(s) are within the region -306 to -97 or are downstream of -96.

The locations of conserved and repeated sequences within the 5' region of the Psa-v gene constructs are shown in Figures 2.1 and 2.2b. The seed expression region between positions -306 and -97, which has at least one important regulatory element, contains vicilin box II and half of vicilin box I, including the octanucleotide and CACA motifs. This region also includes a short palindromic sequence. The region downstream of position -96 contains the CATGCATG, CAAT and TATA elements.
Figure 2.4
Western blot analyses of *Psa-v* gene expression in different organs of tobacco plants transformed with 5' deletion constructs.

Western blots showing levels of vicilin in (A) leaves, (B) stems, (C) roots and (D) mature seed. Lanes -2500, -1476, -1042, -773, -424, -306 and -96 refer to the 5' endpoints of the deletion constructs. Control lanes C contain protein from tobacco transformed with the empty binary vector (pGA492). Protein extracts were immuno-purified and then each lane was loaded with the equivalent of 0.5 mg of total soluble protein.

In blots A, B and C, each lane contains extract from a representative plant from each 5' deletion construct. In blot A, the TE lane contains 1.5 µg of pea total extract, which is equivalent to 600 ng of vicilin. In blots B and C, lanes (+) contain the equivalent of 100 ng of vicilin protein, extracted from transformed tobacco seed.

In blot D, each lane was loaded with protein from a pooled sample of mature seed extracts from 6-9 plants transformed with each particular construct. The pooling was done to indicate the average level of expression conferred by each construct.

Numbers on the vertical axes show the approximate $M_r \times 10^{-3}$ of convicilin (68), vicilin (50) and the major proteolytic vicilin products are indicated on the vertical axes.
2.3.3 Regulation

The expression of the gene is regulated by a combination of cis-acting elements. The results presented here indicate that the pattern of expression is influenced by the presence of a 2500-bp fragment upstream of the transcription start site (Gamp et al., 1998). The presence of this fragment is necessary for the regulation of expression, as indicated by the absence of expression in plants lacking this fragment. To determine the role of the 2500-bp fragment, we performed deletion analysis of the region -2500 to -1476 and -1042 to -773. The results indicate that the 2500-bp fragment is necessary for the regulation of expression, as evidenced by the absence of expression in plants lacking this fragment.

A. Leaf

-2500 -1476 -1042 -773 -424 -306 -96 C C

B. Stem

-2500 -1476 -1042 -773 -424 -306 -96 C C

C. Root

-2500 -1476 -1042 -773 -424 -306 -96 C C

D. Seed

-2500 -1476 -1042 -773 -424 -306 -96 C C

The results from the deletion analysis are consistent with the hypothesis that the 2500-bp fragment contains cis-acting elements necessary for the regulation of expression. The presence of these elements is necessary for the regulation of expression, as evidenced by the absence of expression in plants lacking this fragment.
2.3.3 Temporal regulation

Transcription of the intact \textit{Psa-v} gene in tobacco is regulated in the same temporal fashion as it is in peas (Higgins \textit{et al.}, 1988). The results presented here indicated that the DNA region that regulates the temporal expression of \textit{Psa-v} is downstream of position -306. The patterns of vicilin accumulation in developing transgenic tobacco transformed with constructs \(\Delta-2500\) and \(\Delta-306\) were compared by protein gel blot assays to determine whether deletion of 2200 bp of the upstream sequence influenced the temporal regulation of the gene. Seed from tobacco transformed with \(\Delta-2500\) Figure 2.5, (Higgins \textit{et al.}, 1988), contained detectable levels of vicilin (Mr ~50,000), even in the youngest seed, harvested 9 days after flowering. The level of vicilin increased and then peaked (as a proportion of total seed protein) at 17 days after flowering. Although the amount of vicilin in seed containing the \(\Delta-306\) construct, was much less than that in the \(\Delta-2500\) plant, the patterns of accumulation were very similar (Figure 2.5). Thus, the major element(s) that regulates the temporal expression of \textit{Psa-v} is within the \(\Delta-306\) construct.

2.3.4 Delineation of quantitative \textit{cis}-acting elements

Although the seed-specific and temporal regulatory element(s) appeared to be downstream of position -306, a comparison of the amount of vicilin expressed by different constructs (for example, see Figures 2.4, 2.5) indicated that there were positive, quantitative regulatory elements upstream of this position. Densitometry of Western blots was used to determine the concentration of vicilin in the mature seed of each plant transformed by the 5' deletion gene constructs (Figure 2.6).

There were no significant differences in the vicilin concentrations of plants transformed with gene constructs \(\Delta-2500\), \(\Delta-1476\), \(\Delta-1042\) and \(\Delta-773\) (analysis of variance \(f = 1.125\), \(p>0.25\)) which indicates that there are no major elements between positions -773 and -2500 that modulate the expression of the gene.

In contrast, gene constructs \(\Delta-424\), \(\Delta-306\) and \(\Delta-96\), expressed significantly (\(p<0.05\)) less vicilin; both compared to \(\Delta-773\) (and constructs \(\Delta-2500\), \(\Delta-1476\) and \(\Delta-1042\)), and compared with each other. The deletion of region -773 to -425 caused an 80% decrease in expression. Further deletion, of the region from -424 to -307, reduced the expression to about 10%. No vicilin was
Figure 2.5
Temporal regulation of Psa-v gene expression in developing seed of tobacco transformed with 5' deletion constructs.

The pattern of accumulation of vicilin (and related polypeptides) during seed developments in the transgenic plants Δ-306-G and Δ-2500-B3 is shown. The vicilin levels were determined by densitometric analysis of Western blots. Lanes containing protein extracts from Δ-306-G contained the equivalent of 0.5 mg of total soluble protein, while Δ-2500-B3 extracts were assayed with both 0.5 mg/lane and 0.033 mg/lane, to avoid comparison of Western blots with very different vicilin levels.
The stepwise reduction of expression between constructs A-779, A-Bla, A-206 and A-995, even when 25% more protein was loaded into the gels in some cases, is not shown.

2.4 Discussion

2.4.1 De novo synthesis and temporal regulation

This analysis reveals that the results of the classical assay are de novo specific and temporal regulation. This is in agreement with several other studies, which have shown that the regulatory elements controlling seed-specific gene expression are contained within the initial 306 bp of the 5' region, upstream of the transcription start site (12). This has been demonstrated for genes coding for both 7S (α-) and δ-cyoglobin (Chan et al., 1993) and β-conglycinin (Robertson et al., 1991; Burrow et al., 1990) and α-L proteins (Jones, Robertson et al., 1990) and also to α 22
detected (that is; less than 10 ng/mg total protein) in the seed of any of the Δ-96 plants. Furthermore, vicilin was not detected in the seeds of plants Δ-96B and Δ-96G, even when 25 times more protein was loaded into the gel lanes (data not shown).

The step-wise reduction of expression between constructs Δ-773, Δ-424, Δ-306 and Δ-96, correlated with the loss of the AT-rich repeats (Figure 2.6). Constructs Δ-2500, Δ-1476, Δ-1042 and Δ-773, all expressed high levels of vicilin in transgenic tobacco seed, and contain both AT-rich repeats; while construct Δ-424 contains (most of) one repeat and construct Δ-306 has none. Hence the AT-rich sequences appear to contain the cis-acting elements that are required for high levels of expression of Psa-v.

2.4 Discussion

2.4.1 Seed-specific and temporal regulatory elements within Psa-v.

This chapter describes the results of experiments that were designed to locate regulatory elements that control the seed-specific, temporal and quantitative expression of a vicilin gene, Psa-v. This was done by analysing transgenic tobacco transformed with a number of Psa-v gene constructs that contained different lengths of the 5' region. Three major regulatory regions were identified; two regions contain elements that regulate the level of expression and one region contains element(s) that control seed-specific and temporal regulation.

SEED-SPECIFIC AND TEMPORAL ELEMENTS ARE LOCATED DOWNSTREAM OF POSITION -306

Deletion of the 5' region of the Psa-v gene indicated that the regulatory elements that control seed-specific regulation are downstream of position -306. This is in accord with several other studies, which have shown that the regulatory elements controlling seed-specific gene expression are contained within the initial 306 bp of the 5' region, upstream of the transcription start site (Table 1.1). This has been demonstrated for genes coding for both 7S (α',β–conglycinin, Chen et al., 1986; and β-phaseolin Bustos et al., 1991; Burow et al., 1992) and 11S proteins (glycinin, Nielsen et al., 1989) and also for a 2S
Figure 2.6  
Concentration of vicilin in seed of tobacco transformed with 5' deletion constructs.

Average vicilin concentration (ng vicilin /mg total protein) +/- standard error (s.e.) in mature seed from tobacco transformed with Psa-v constructs with 5' deletions at positions -2500, -1476, -1042, -773, -424, -306 and -96. Seed from six to nine transformed plants were assayed for each construct. The concentration of vicilin for each plant was determined by densitometric analysis of Western blots. No vicilin (detection limit; <10 ng vicilin/mg total protein) was detected in any of the plants transformed with construct A-96. Arrows and markers within the diagrams of the 5' deletion constructs represent putative cis-acting elements as described in Figures 2.1 and 2.2.B.
### Psa-v 5' deletion mutants

<table>
<thead>
<tr>
<th>AT-rich</th>
<th>1-1</th>
<th>vicilin [ng/mg total protein]</th>
</tr>
</thead>
<tbody>
<tr>
<td>-2500</td>
<td>![Diagram]</td>
<td>( x \pm s.e. )</td>
</tr>
<tr>
<td>-1476</td>
<td>![Diagram]</td>
<td>( 470 \pm 110 ) 7</td>
</tr>
<tr>
<td>-1042</td>
<td>![Diagram]</td>
<td>( 810 \pm 170 ) 7</td>
</tr>
<tr>
<td>-773</td>
<td>![Diagram]</td>
<td>( 750 \pm 140 ) 9</td>
</tr>
<tr>
<td>-424</td>
<td>![Diagram]</td>
<td>( 710 \pm 100 ) 8</td>
</tr>
<tr>
<td>-306</td>
<td>![Diagram]</td>
<td>( 150 \pm 40 ) 7</td>
</tr>
<tr>
<td>-96</td>
<td>![Diagram]</td>
<td>( 65 \pm 15 ) 7</td>
</tr>
</tbody>
</table>

*Vicilin was not detected in the absence of a seed specific regulatory activity of the Psa-v promoter.*
albumin gene (napin, Radke et al., 1988). Many lectin genes are highly expressed in seeds, but are also expressed in other tissues. Such genes also contain the regulatory elements responsible for tissue-specific expression within the proximal part of the promoter (soybean, Okamuro et al., 1986; *Dilichos biflorus*, Harada et al., 1990; common bean, Riggs et al., 1989).

Similarly, the location of seed-specific regulatory elements in cereal genes have been shown to be near the transcription start site (Colot et al., 1987; Halford et al., 1989; Robert et al., 1989; Matzke et al., 1990; Entwistle et al., 1991; Quattrocchio et al., 1992; Thomas and Flavell, 1990; Takaiwa et al., 1991).

Vicilin was not detected in the roots, leaves or stems of plants transformed with any of the *Psa-v* 5' deletion constructs. This could be because the genes were not transcribed, the transcripts were unstable, or because the protein was degraded rapidly. Plants transformed with chimeric genes consisting of 7S coding regions ligated to the CaMV 35S promoter, express detectable levels of 7S mRNA and protein in leaves and other non-seed organs, (Lawton et al., 1987; Higgins and Spencer, 1991; Wandelt et al., 1992; Bagga et al., 1992) although 7S proteins are less stable in leaf tissue (Bagga et al., 1992; Wandelt et al., 1992). Thus, it seems unlikely that the lack of vicilin in plants transformed with the *Psa-v* 5' deletion constructs can be attributed wholly to instability of vicilin, or its transcript in these tissues. The expression of a gene construct containing the *Psa-v* promoter ligated to a chloramphenicol acetyl transferase (CAT) reporter gene was also found to be undetectable in the leaves, stems or roots of transformed tobacco (Appendix II). Since the CAT enzyme is relatively stable in these non-seed tissues, these results support the proposal that the absence of protein was primarily because the *Psa-v* gene constructs were not transcribed.

In addition to being seed-specific, the expression of vicilin genes is restricted to relatively early stages of seed development. The expression of vicilin in the \(\Delta-2500\) *Psa-v* construct and the \(\Delta-306\) construct was assayed at different stages of seed development to assess whether temporal regulation could be distinguished from seed-specific regulation. The pattern of expression of \(\Delta-306\) in developing tobacco seed was similar to that of the \(\Delta-2500\), indicating that the essential temporal regulatory element(s) are downstream of position -306.
As in this study, most other studies have been unable to distinguish between the regions controlling temporal regulation from those containing the seed-specific regulatory elements (Colot et al., 1987; Chen et al., 1986; Marris et al., 1988; Shirsat et al., 1989; Murai and Kawagoe, 1991; Rerie et al., 1991; Takaiwa et al., 1991; Fujiwara and Beachy, 1992). This raises the question as to whether seed-specific and temporal regulation are regulated by the same DNA elements, or whether there are separate elements (see Chapter 1).

Detailed analyses of the 5' region of Pvu-β, a 7S gene coding for phaseolin in soybean, however, have located separate regions that are involved with temporal and seed-specific regulation. Two DNA regions associated solely with temporal regulation have been identified (Bustos et al., 1991). The region containing the major temporal element, however, also represses expression in stem and root tissue (Burow et al., 1992). This indicates that some (although probably not all) aspects of the temporal regulation of seed storage protein genes may be activated and repressed by elements that do not actively regulate seed-specific expression. Hence, temporal elements may moderate expression during seed development but be distinct from other elements that moderate expression in non-seed tissues.

Further studies will be required to determine whether the temporal element(s) within Ps-a-v act as separate domains, or as part of the seed-specific element(s).

Many seed-specific sequences are in the region of Ps-a-v promoter that confers seed-specific expression

Many promoters and enhancers that control transcription in mammalian cells are composed of multiple interacting regulatory elements (Dynan, 1989; Mitchell and Tjian, 1989) each of which may affect the structure of the DNA or be specifically recognised by one or more trans-acting proteins. Light regulated genes, for example, are regulated by a complex array of short regulatory sequences (Gilmartin et al., 1990; Kuhlemeyer et al., 1987) as is the 35S gene from CaMV, which has seven interacting domains within a 351 bp stretch of the promoter (Benfey and Chua, 1990). Since these regulatory regions are often conserved and replicated in the promoters of similarly regulated genes, sequence comparisons within and among 5' regions have been used to identify putative cis-acting sequences.
Seed-specific and temporal regulation was conferred by the element(s) downstream of -306 of \textit{Psa-v}. A number sequences that appear to be conserved between different seed-specific genes, and so might contain elements that regulate gene expression (see Figures 2.1 and 2.2b and Chapter 1, for review) are present within the DNA sequence between -306/+1 of the \textit{Psa-v} gene. These sequences include a CATGCATG element, also called an RY repeat, (Dickinson \textit{et al.}, 1988), a CACA element (Goldberg, 1986) an octanucleotide sequence (Rerie \textit{et al.}, 1992) and vicilin box II (Newbiggin, 1988).

The CACA element and octanucleotide sequence are within a larger conserved sequence. This sequence, vicilin box I, is highly conserved in 7S genes from legumes (Bown \textit{et al.}, 1988; Gatehouse \textit{et al.}, 1986; Higgins \textit{et al.}, 1988; Rerie \textit{et al.}, 1992; Table 1.3), but it is not found in the vicilin genes for cotton (see Figure 1.2) or cocoa (McHenry and Fritz, 1992), which suggests that this motif may be conserved only in the genes for legume 7S proteins.

The location and possible functions of the conserved DNA elements in the two subsections of the seed specific regions are discussed below. At least one regulatory element is located within -306 to -97 of \textit{Psa-v}, since deletion of this sequence reduced expression below detectable levels (Figure 2.6). The sequence -306/-97 was required for detectable levels of expression in the seed and so was termed the seed-expression region. Most of the conserved elements are present within this region, including vicilin box I, containing the octanucleotide and the CACA motifs, and vicilin box II, but the CATGCATG motif is downstream of position -96.

The most likely conserved sequence to contain a seed-specific regulatory element of those located within -306 to -97 is the octanucleotide box. The octanucleotide box appears to be present within promoters of many genes that are highly expressed in (dicotyledonous) seeds, and is absent from the promoters of genes that are primarily expressed in non-seed tissues (Table 1.2).

Functional analyses of genes that are highly expressed in seeds indirectly support the proposal that the octanucleotide box may have a functional role. The octanucleotide sequence (and two copies of the CACA motif) are present within the region -227 to -106 of the gene \textit{Pvu-\beta}, which codes for phaseolin, a 7S protein from common bean (\textit{Phaseolus vulgaris}). This region contains an element that contributes to seed-specific expression by suppressing expression
within stem and root tissue (Burow et al., 1992). Secondly, the octanucleotide sequence is present within the 5' region of a lectin allele that is highly expressed in common bean seeds. The motif is absent in another, almost identical allele, that is expressed at very low levels in seeds (Voelker et al., 1987), which indicates that the octanucleotide box might also contribute to high-level expression. Thirdly, a gene encoding α',β-conglycinin, Cgy1, also contains an octanucleotide box at about -130 (Figure 1.3). Insertion of the region -245/-57 of Cgy1, into the CaMV 35S promoter, was shown confer 40-fold enhancement of seed-specific expression in transgenic tobacco. It was previously thought that this enhancement was mediated by multiple ADCCCA copies (Chen et al., 1988; D=A,C,G), but mutational analysis has shown that the ADCCCA sites can be deleted without affecting expression (Fujiwara and Beachy, 1992)), and so other conserved sequences present within this region, such as the octanucleotide CATGCATG and CACA elements (see below), may have contributed to this seed-specific enhancement of expression.

The CACA element is also present at a significantly high frequency within the promoter regions of seed storage protein genes from both cereals and legumes (Table 1.2), and so might act as a seed-specific regulatory element. Although at least one report indicates that the CACA element may be involved in expression of seed storage protein genes (Aryan et al., 1991), the prevalence of this motif in other genes that are expressed at later stages of development and during germination (see column "other", Table 1.2, and also Appendix I) suggests that it may be functional during other stages of seed development. The motif is by no means restricted to the promoters of genes expressed highly in seeds, however, and so might act as a general enhancement element (Table 1.2).

Vicilin box II, unlike the CACA element, is highly gene-specific. It is present only in the 5' regions of vicilin genes (Table 1.2). Deletion analysis of another 7S gene, CGy1, however, showed that sequences between -457 and -257 could be deleted without significantly affecting gene expression in transgenic petunia (Chen et al., 1986). This region contained vicilin box II, which suggests that this conserved sequence may not have an essential role in regulating gene expression.
The CATGCATG motif was not sufficient for detectable gene expression

The gene construct Δ-96 was not expressed, even though, in addition to the CAAT and TATA boxes, it contained the only copy of the CATGCATG element (sequence tATGCATG) in the Psa-v promoter. A second CATGCATG-like element, ggTGCATG, is also located (at -68) in this region. CATGCATG is a very highly conserved motif which is present in the 5' region of many genes that are highly expressed in seeds, especially the globulin genes of legumes (Dickinson et al., 1988; Table 1.2).

The results reported here are consistent, however, with those obtained for other seed-storage protein genes (Chen et al., 1986; Baumlein et al., 1991; Rerie et al., 1991). A 5' deletion analysis of a legumin gene from pea, LegA1, showed that -237 bp of the 5' flanking sequence was insufficient for expression of the legumin gene in transgenic tobacco, even though it contained a CATGCATG motif (as part of the conserved legumin box, Rerie et al., 1991). Similarly, deletion analysis of the LeB4 promoter (Baumlein et al., 1991) and the CGy1 promoter (Chen et al., 1986) demonstrated that the CATGCATG sequence was not sufficient for high-level expression in seeds.

Site-specific mutagenesis, however, has shown that the CATGCATG motif is vital for seed specific regulation. Destruction of the CATGCATG sequence by a 6 bp deletion within the otherwise intact promoter of the faba bean (Vicia faba) legumin gene LeB4 drastically reduced expression in seeds and resulted in low level expression in leaves (Baumlein et al., 1992). Similarly, the mutation of the two CATGCAT/c sequences within the promoter of the soybean α',β-conglycinin (7S) gene, CGy1, reduced the level of gene expression in seeds (Fujiwara and Beachy, 1992; Chamberland et al., 1992) and led to expression in leaves (Fujiwara and Beachy, 1992). Internal deletion of the 7 bp of a CATGCAT sequence in a glycinin gene, Gy2, from soybean (Glycine max) also caused a reduction in gene expression (Lelievre et al., 1992).

Thus, although the CATGCATG motif has been shown to be essential for high-level, seed-specific regulation of genes encoding for seed storage proteins from faba bean and soybean, it is not sufficient, and interaction with additional upstream elements is required.
Does the sequence -306/-229 contain a regulatory element?

Although CATGCATG, vicilin box I and vicilin box II are present in the seed-specific region of Psa-v, previous results had shown that the region between -228/-49, which contains these conserved motifs, was not sufficient for expression (Newbigin, 1988 and Appendix II). This suggested that additional regulatory elements, upstream of position -229, were also required for expression. This has been partially confirmed by this study, as the sequences between -773 and -307 were required for full, high level expression of Psa-v (Figure 2.6). Low level expression, however, was detected with construct Δ-306. Thus, there may be a positive regulatory element in the 78 bp sequence between -306 to -229, perhaps within the small interrupted palindrome.

Alternatively, the sensitivity of the CAT assay, or the stability of CAT in seeds, may be less than that for vicilin. The background level of endogenous CAT activity in tobacco seed was quite high, but it was estimated that expression levels of 12 % or more (of that obtained by the full length promoter) could be discerned (Newbigin, 1988). Expression of Δ-306 ranged between 3-16% of that driven by the full-length 5' region, and so, if pEN12 was expressed at a similar level, one might have expected to observe activity in the seed of at least one of the six plants transformed with pEN12.

Since many sequences that act as regulatory elements in eukaryotic genes are bound by nuclear proteins (Maniatis et al., 1987; Mitchell and Tjian, 1989; Peterson and Baichwal, 1993), DNA-protein binding studies can be used to locate sites that may contain cis-acting elements. Analysis of the binding of nuclear proteins to the conserved motifs, AT-rich repeats and other regions of the Psa-v promoter are described in Chapter 3. DNA-protein binding studies to other seed-storage protein genes are also reviewed in Chapter 1. Functional analysis of the regions that were bound by nuclear proteins and/or conserved DNA sequences, to find out which, if any, of these sequences act as seed-specific, temporal or quantitative regulatory elements, is described in Chapter 4.

In order to resolve which, if any of the conserved motifs act as regulatory elements within the seed-specific region of the Psa-v gene, and whether the region between -306/-229 has a functional role, a set of constructs was made containing internal deletions within the region -306 to +1. Analysis of the expression of these constructs is described in Chapter 4.
2.4.2 Quantitative elements in \textit{Psa-v}

\textbf{THE QUANTITATIVE ELEMENTS IN \textit{Psa-v} CONTAIN AT-RICH REPEATS}

Two regions containing major and minor quantitative elements were identified within the 5' region of \textit{Psa-v}. The regions, -773/-425 and -424/-307 respectively, were searched to find repeated or conserved sequences that might contain these regulatory elements. Both regions contain almost identical AT-rich repeats which are 93 bp long with 93% identity (Figure 2.1; original sequence comparison identified 55 bp repeats within these sequences; Higgins \textit{et al.}, 1988). They extend from -550 to -458 and -462 to -370, and so overlap slightly.

There are at least two explanations why major quantitative region, -773/-424, had a greater effect upon the enhancement of gene expression than the minor quantitative region, -424/-307. The major quantitative region could contain additional regulatory elements; construct \(\Delta-424\) contains only 56 bp of the proximal repeat, and so other elements might be present in the remainder of the AT-rich repeats, between -550/-513 and -457/-425, or in other DNA sequences within -773/-551. Alternatively, the presence of a second AT-rich element (between-512/-458) might have a synergistic effect, rather than an additive effect upon expression. Synergistic interactions between copies of identical elements have been observed in other genes expressed in plants including the 35S enhancer (Kay \textit{et al.}, 1987), ARE (Olive \textit{et al.}, 1990). For example, the soybean heat shock gene Gmhsp17.5E also contains an AT-rich quantitative element; the presence of multiple copies of this 33 bp element had a synergistic effect upon induction of gene expression, such that 2, 3 and 5 copies gave 10, 15 and 24 fold induction (Czarnecka \textit{et al.}, 1992).

\textbf{AT-RICH QUANTITATIVE ELEMENTS DO NOT CONFER SEED-SPECIFIC EXPRESSION}

Deletion of the AT-rich quantitative regions within the 5' region of \textit{Psa-v} did not affect the temporal or seed-specific expression of the gene constructs, which indicates that the quantitative elements may have little or no role in actively maintaining seed-specific regulation. This is supported by results from experiments that studied the effect of AT-rich quantitative elements in the absence of the proximal region of seed-specific gene promoter sequences. Fusion of the AT-rich region from \textit{Pvu-\(\beta\)}, which codes for phaseolin, to a GUS gene containing a minimal 35S promoter (-89/+6) yielded strong GUS
expression in several tissues, especially roots (Bustos et al., 1989). AT-rich fragments from a sunflower genes, (regions -1527 to -75 and -739 to -75 of HaG3A), which codes for the 11 S protein helianthinin, also confer expression upon non-seed tissues, especially roots (Bogue et al., 1990). Since the region of the 35S promoter that was used (domain A) confers low level expression principally in roots and embryonic tissue destined to become roots (Benfey et al., 1989), the AT-rich regions may enhance the expression of this A domain, rather than containing intrinsic root-expression elements.

The proposal that these AT-rich regions may be essentially neutral quantitative elements (that do not affect organ-specific or temporal regulation) is also supported by experiments which analysed the expression of chimeric genes with combinations of the 5' regions of legumin and vicilin genes from pea. Ligation of a putative quantitative element from LegA1 to an otherwise non-expressing chimeric gene, containing -228/-49 of Psa-v, resulted in seed-specific expression in transgenic tobacco (Appendix II). Similarly, an AT-rich element from Gmhsp17.5E enhanced anaerobically induced transcription of a chimeric promoter containing a truncated (-140) Adh1 promoter. No transcription was detected in unstressed tissue, indicating that the enhancement of expression by these AT-rich sequences was restricted to tissues in which the other promoter elements were active (Czarnecka et al., 1992).

**AT-RICH ELEMENTS ARE COMMON TO MANY GENES**

AT-rich sequences have been found in positive regulatory elements in 5' regions of many other seed-expressed genes, usually located -1000 to -500 bp upstream of the start of transcription. Examples of AT-rich quantitative regions include those in genes coding for 11S proteins (Bogue et al., 1990; Bustos et al., 1989; Jordano et al., 1989; Rerie, 1989; Shirsat et al., 1989; Shirsat et al., 1991; Rerie et al., 1991; Lelievre et al., 1992; Baumlein et al., 1992; Jordano et al., 1989), and lectins (Lindstrom et al., 1990, Riggs et al., 1989), prolamines (Thompson et al., 1990), and other seed-expressed genes (Marcotte et al., 1989; Roth et al., 1991; Goupil et al., 1992). This indicates that AT-rich sequences, such as those in Psa-v, may be members of common, loosely conserved set of elements that have a similar function in the regulation of seed-expressed genes.
AT-rich sequences are also present in the quantitative elements of many other plant genes. These include genes expressed in nodules (Szabados et al., 1990), those responding to light (Castresana et al., 1988; Datta and Cashmore, 1989; Cushman and Bohnert, 1992), heat shock (Czarnecka et al., 1992; Rieping and Schoffl, 1992) and other stresses (Harrison et al., 1991). Given the presence of AT-rich regions in quantitative regions of genes that are regulated in such different ways, this suggests that AT-rich motifs may represent a family of general quantitative elements that are present in all types of genes, not just those expressed in seed.

To summarise; three regulatory regions were identified in the 5' region of the \textit{Psa-v} gene. The regulatory region(s) controlling seed-specific and temporal regulation were downstream of position -306 and contained the conserved sequences, vicilin boxes I and II, and a CATGCATG element. Deletion of region -306 to -97, which included both vicilin boxes, resulted in loss of detectable expression, indicating the presence of at least one important regulatory element. Two quantitative elements were also identified; between -773 to -425 and -424 to -307. Both contained almost identical AT-rich repeats; which resemble AT-rich sequences found in positive elements in other seed- and non-seed-specific promoters. Further analysis is required to determine which sequences are recognised by transcription factors and to precisely define the role of particular elements in regulating high-level, seed-specific expression of the \textit{Psa-v} gene.
CHAPTER 3. DNA-PROTEIN BINDING IN THE 5' REGION OF PSA-V

3.1 Introduction

Transcriptional regulation is modulated by the binding of sequence-specific transcription factors to their target sequences in the genes, and the subsequent interactions between these transcription factors and other proteins, including members of the transcription initiation complex (see Chapter 1 for review). Several putative plant transcription factors that bind to specific DNA sequences, in genes associated with light regulation or with developmental regulation in flowers, leaves or trichomes, have been isolated (reviewed in Katagiri and Chua, 1992).

Transcription factors (or putative transcription factors) that regulate the expression of genes in seeds have been also been identified, although all of these are from cereals, rather than legumes. Opaque-2 is a bZIP protein from maize that is required for high level transcription of genes encoding 22 kDa zeins. Activation of transcription of these genes is mediated by the binding of this transcription factor to a conserved sequence in the promoters of these zein genes (Hartings et al., 1989; Schmidt et al., 1990; Lohmer et al., 1991).

Likewise, the embryo maturation (Em) gene from wheat, which is expressed during the later stages of seed development, contains a site that is bound by a bZIP protein, EmBP-1. This site, and presumably the binding of EmBP-1, is required for ABA-induced gene expression (Marcotte et al., 1989; Guiltinan et al., 1990; Marcotte et al., 1992).

Specific DNA sequences in the 5' regions of legume seed storage genes that are bound by nuclear proteins which may be transcription factors, have also been identified. The majority of binding sites that have been identified are in AT-rich regions of DNA. These regions, however, are associated with high level expression, rather than with seed-specific or temporally-regulated gene expression (Weising and Kahl, 1991; and Table 1.1, for review). For example, the genes encoding phaseolin, (-682 to -628, Bustos et al., 1989), and helianthinin, (-725 to -322, Jordano et al., 1989; Bogue et al., 1990) contain AT-rich DNA sequences that are bound by nuclear proteins. When these binding sites were ligated to a truncated CaMV 35S promoter (-90), they conferred high level expression upon a GUS gene in the roots (Bustos et al., 1989; Jordano et al., 1989; Bogue et al., 1990). Thus, although these AT-rich binding regions contain cis-acting elements that can activate gene expression,
they are not directly associated with regulating seed-specific expression.

Some DNA-binding proteins have been observed, however, binding in regions associated with temporal and/or seed-specific regulation of two genes for 7S proteins. Multiple binding sites for nuclear proteins (Allen et al., 1989; Kawagoe and Murai, 1992) have been identified in the seed-specific and temporal elements in the phaseolin and \( \alpha',\beta \)-conglycinin genes, \( Pvu-\beta \) and \( Gma-\alpha' \) (Chen et al., 1986; Chen et al., 1988; Burow et al., 1992). The 5' region of the vicilin gene \( Psa-v \), encoding a 7S protein in pea, is similar to the \( Pvu-\beta \) and \( Gma-\alpha' \) genes. The first 300 bp of these genes share 50% and 57% identity with \( Psa-v \), respectively (Chee and Slightam, 1991), and these promoter regions each contain the conserved vicilin box I sequence. (see Figure 1.4). Also, both \( Psa-v \) and \( Gma-\alpha' \) contain a copy of vicilin box II (Newbigin, 1988; and Appendix I). In addition, \( Psa-v \), like these other 7S genes, contains AT-rich regions upstream of the proximal promoter region (Slightom et al., 1983; Doyle et al., 1986; Higgins et al., 1988). As well as sharing similar 5' regions, all three genes are expressed in a seed-specific and temporally-regulated manner (reviewed in Rerie et al., 1992), and so these genes may be regulated by homologous transcription factors. Hence the 5' region of \( Psa-v \) may contain similar DNA binding sites to those in the \( Pvu-\beta \) and \( Gma-\alpha' \) genes.

When the 5' region of the \( Psa-v \) gene was progressively deleted, a regulatory region (termed the seed-specific region) essential for seed-specific and temporal expression was identified downstream of position -306. Major and minor quantitative elements that regulate the level of expression were located between -773/-425 and -424/-307 (Chapter 2). Deletion of the region from -306 to -97 resulted in loss of detectable gene expression, indicating the presence of at least one major regulatory element within, or overlapping, this region. Thus, the sequence between -306 to -97 has been termed the seed-expression region. This does not exclude the possibility, however, that -307/-97 might contain only quantitative elements, and that the seed-specific and temporal elements may be located downstream of position -96 (see Chapter 4).

In this study, gel-shift assays were used to determine DNA-binding sites for nuclear proteins in the regulatory regions of the \( Psa-v \) gene; in particular in the seed-expression region between positions -306 to -97. The binding profiles of nuclear extracts from different stages of development were analysed to determine whether the DNA-protein binding activity varied substantially during seed development. Nuclear extracts from leaves were also tested, to
determine whether the DNA-binding proteins were present in other tissues, and whether the nuclei of this vegetative organ contained factors which might repress the binding of the seed DNA-binding proteins.

Four seed-specific nuclear proteins were identified that bound to specific regions within -773 to +1 of the Psa-v gene. The binding activities of three of these, TEMP-1, TEMP-2 and TEMP-3, were temporally-regulated, strongly peaking between 11 and 16 days after flowering (DAF), prior to the major peak in vicilin mRNA levels. Since TEMP-1 bound within the seed-expression region, and TEMP-2 and -3 bound within the remainder of the seed-specific/temporal region, these proteins may have a role in temporal and/or seed-specific regulation. The fourth nuclear protein identified, AT-2, bound to an AT-rich consensus sequence that was present in both the major and minor quantitative elements, indicating that it may have a role in enhancing transcriptional activation. Leaf extracts were found to contain a factor which specifically abolished the DNA-binding activity of TEMP-1 and AT-2 (TEMP-2 and -3 not tested), and this suggests that the DNA-binding activity of these proteins may be regulated by protein modification, or by binding with an inhibitory protein.

3.2 Materials and Methods

3.2.1 Preparation of crude nuclear extracts

Crude nuclear extracts were isolated from seeds and leaves from peas (Pisum sativum L. cv. Greenfeast) grown in the glasshouse at 15-22°C. Extracts were made from seeds at six stages of development. The stages (based on fresh weight in mg/pair of cotyledons) were 50-100, 100-200, 200-300, 300-400 and 400-500 which approximated 8-10, 11-12, 13-15, 16-17, and 18-20 DAF, respectively. Fresh leaves were taken from younger plants, that were growing vigorously prior to flowering. These were incubated in diethyl ether in order to strip the cuticle layer, then washed three times prior to extraction in nuclei wash buffer, (100 mM mannitol, 30 mM HEPES buffer pH 7.5, 10 mM MgCl₂, 5 mM β-mercaptoethanol, at 4°C; Watson and Thompson, 1986).
Nuclei were isolated essentially by method II of Luthe and Quatrano (1980). Briefly, 5 g of cotyledons (frozen at -80°C), or leaves were homogenised in 12.5 ml of extraction buffer, containing 1000 mM hexylene glycol, 10 mM Pipes buffer pH 7.0, 10 mM MgCl₂, 15 mM β-mercaptoethanol, at 4°C (Watson and Thompson, 1986). The homogenate was centrifuged at 1000 x g for 10 min and the pellet was resuspended in 2.5 ml of nuclei wash buffer. The extract was centrifuged at 1000 x g for 5 min, and the pellet was resuspended in 0.5 ml of sonication buffer (20 mM HEPES pH 7.5, 420 mM KCl, 25% glycerol (v/v), 1.5 mM MgCl₂, 0.2 mM EDTA, 5 mM β-mercaptoethanol).

The nuclei were lysed by incubation on ice for 1 hour, followed by sonication with two 5 second bursts at approximately 50 W, using a Branson sonicator. The cellular debris was pelleted by centrifugation at 10 000 x g for 10 min. The supernatant was dialysed for 5-8 hours against dialysis buffer (20 mM HEPES pH 7.5, 50 mM KCl, 0.2 mM EDTA, 20 % (v/v) glycerol, 5 mM β-mercaptoethanol) and then centrifuged at 10 000 x g for 8 min. The supernatant was frozen in liquid nitrogen and stored at -80°C. The protein content of the nuclear extract was determined by the Bradford (1976) method, using bovine serum albumin (BSA) as a protein standard.

Maize nuclear extract was prepared from a Zea mays cv. Black Mexican Sweet XII-II suspension cell line supplied by Chourey and Zurawski (1981), as described by Tokuhisa et al. (1990). The extract was kindly donated by Mark Olive, Jeff Ellis and Jim Tokuhisa. The E. coli extract which contained the DNA-binding domain from EBNA-1 (an Epstein-Barr virus-encoded DNA-binding protein) from a Pharmacia (Australia) Sure-Track footprinting kit, was generously donated by Peggy Horn.

### 3.2.2 Cloning of DNA fragments

Five fragments from the 5' region of Psa-v were subcloned into pUC118, for easy labelling and isolation of individual DNA fragments (Figure 3.1). Plasmid pCW70, containing the Psa-v gene from -773 to +2699, was cut at the Xba I site in the polylinker 5' to the gene, and at an Acc I site at position -423. The -773/-423 fragment was subcloned into the Xba I and Acc I sites of pUC118 to generate plasmid pQ9. Similarly, plasmids pCW74 and pCW71 containing the 5' deletions Δ-424 and Δ-306 (Chapter 2) were cut at the Sal I site (in the polylinker) and the Hae III site (position -231) to generate fragments -424/-231...
and -306/-231, respectively. These were subcloned into pUC118 at the Sal I and Sma I sites, to create plasmids pQ6 and pQ7. Plasmid pQ8 was made by the ligation of the Hae III - Hae III fragment, from pCW74 -230/-78, into the Sma I site of pUC118. Plasmid pQ18 was made by the ligation of the Nde I - Cla I fragment from pCW70, containing sequences between -96 to +1, into the Xba I and Acc I sites of pUC118.

The -666/-236 fragment of the pea legumin gene Leg A1 (Rerie et al., 1990) was excised by digestion with Ssp I and Nco I and then subcloned into the Sma I site in pUC8 to generate plasmid pQ3. All cloning procedures were essentially as described in Maniatis et al. (1982) and plasmids were verified by DNA sequencing using a Taq DyeDeoxy terminator cycle sequencing kit (Applied Biosystems, Australia).

3.2.3 DNA fragments used for probes and competitor DNA

Plasmids pQ9 and pQ18 (Figure 3.1C) were digested with Eco RI and Hind III to generate fragments containing -773/-423 and -96/+1 which were used as probes and competitor DNA. Plasmids pQ6, pQ7 and pQ8 were digested with Xba I and Eco RI; which released fragments containing -424/-231, -306/-231 and -230/-78, respectively.

Gene constructs, with nucleotide substitutions (Σ) within the seed-specific region were used as competitor DNA. Constructs pQ-ΣA-306/-272 and pQ-ΣB-306/-231 contain Psa-v fragments -272/+39 and -231/+39, respectively. The constructs pQ-ΣC-230/-194, pQ-ΣD-198/-175, pQ-ΣE-186/-150, pQ-ΣF-153/-117, pQ-ΣG-129/-96, pQ-ΣH-98/-79, pQ-ΣI-88/-73, pQ-ΣK-61/-4 all contain the region from -306 to +39, with substitution within the flanking regions indicated in the plasmid names (Figure 3.2; see Chapter 4, and Figure 4.1 for details of construction). The fragments containing these mutations were excised by digestion with Hind III, (at the polylinker 5' of -306) and Nco I (+39) site. For brevity, these constructs are referred to below without their pQ-prefixes.
Figure 3.1
The 5' region of Psa-v: Conserved sequences, regulatory regions and fragments used in gel-shift assays.

A. Conserved and repeated sequences in -773 to +1 of the 5' sequence of the Psa-v gene. The location of conserved and repeated DNA sequences within the 5' region of Psa-v are shown; from right to left: TATA box (T), CAAT box (C), cATGCATG element (▲), vicilin box I (overlined, containing a CACA element, ◀ and an octanucleotide box, ▼), vicilin box II (▼) a palindrome (leftrightarrow) and AT-rich repeats (leftrightarrow). Positions are relative to the start site of transcription.

B. Quantitative and seed-specific regulatory elements between -773 and +1 are shown, as defined in Chapter 2. The number of crosses indicates the relative level of expression conferred by these elements. The positions of the seed-specific region (which was sufficient for seed-specific and temporally-regulated expression) and the seed expression region (which was necessary for seed expression and may contain quantitative, seed specific and/or temporal elements) are indicated.

C. Fragments of the 5' region of Psa-v that were used in gel-shift assays were subcloned into pUC118. This resulted in plasmids: pQ9, contains the major quantitative element; pQ6, contains the minor quantitative element and part of the seed-expression region; pQ7, contains region -306 to -231 of the seed-expression region; pQ8, contains region -230 to -78 of the seed-specific region; and pQ18, with -96 to +1 of the seed-specific region.
A. AT-rich repeats

B. ++++++ seed-expression region
quantitative regions seed-specific region

C. Psa-v

pQ9
pQ6
pQ7
pQ8
pQ18
Figure 3.2
Mutations within the 5' sequence of *Psa-v*.

Deletion endpoints for each construct represent the 3' and 5' base pairs flanking the mutated sites. Unmutated bases are represented by dots. The polylinker regions that were used to replace segments of the seed-specific region of *Psa-v* are shown; bases identical to the original sequence are in upper case, other bases are in lower case. Insertions are in superscript and deletions represented by a dash. Numbering is relative to the transcription start site, with a six base pair correction at position -296, in lower case. A small interrupted palindrome and the conserved vicilin box II, octanucleotide, CACA, CATGCATG (mismatches in lower case), CAAT and TATA boxes are highlighted. The 33 bp vicilin box I is underlined.
### Internal Mutations within the Seed-Specific Region of *Psa-v*

<table>
<thead>
<tr>
<th>Construct</th>
<th>Deletion</th>
<th>Palindrome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wild type</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>-306 AAGACT ggaCTTGG TTGTAAGATA CAACACTTGTC TAAAATTTCG TAGTCAATTC ATTAATTCG ATACACATG TCCAAGGCAA TAAATAAGCA TTCCTCCTTTT</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>-230/-194</td>
<td>gacTctA ggaTcccgg gtaCtagag</td>
</tr>
<tr>
<td>C</td>
<td>-231</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Construct</th>
<th>Deletion</th>
<th>Vicilin Box II</th>
<th>Octanucleotide</th>
<th>CACA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wild type</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A,B</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>-230/-194</td>
<td>gactcata ggaCTCTAGA ggatc</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>-198/-175</td>
<td>gAcTCTa GgaCTCTaga ggatc</td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>-186/-150</td>
<td>gAcTCTa tagAggaTCC cctgAtgc attgatgC-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>-153/-117</td>
<td>gACTCTA agTagCaCtCg aCCTAGAtg acGc</td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>-129/-96</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H,I,K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Construct</th>
<th>Deletion</th>
<th>CATGCATG</th>
<th>CAAT</th>
<th>catGCATG</th>
<th>TATA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wild type</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A,B,C</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D,E,F</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>-129/-96</td>
<td>Cgataatt</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H</td>
<td>-98/-79</td>
<td>gACTCTa gaGagAtc-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>-88/-73</td>
<td>gAcTctA gaGagAtc-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>K</td>
<td>-61/-4</td>
<td></td>
<td></td>
<td>cCTg caggTcGACT CtagaggatC</td>
<td></td>
</tr>
</tbody>
</table>
Smaller fragments of the minor quantitative element and the seed-expression region (as shown in Figures 3.15 and 3.16) were made by digestion of the \textit{Psa-v} fragments from plasmids pQ6, pQ7 and pCW74 (Chapter 2). Further digestion of the pQ6 fragment at \textit{MseI} (-399) and \textit{DdeI} (-340, -308) restriction sites produced sub-fragments -424/-400, -424/-341, -399/-341 and -340/-309. The \textit{SalI/HaeIII} fragment of pCW74 from -306 to -231, the \textit{MaeI} fragment of pQ7 from -306 to -261, and the \textit{MaeI/KpnI} fragment of pQ7 from -260 to -231 were also isolated and used as competitor DNA.

Smaller sub-fragments of the -230/-78 fragment of the seed-specific region (as shown in Figure 3.11) were generated by digesting pQ8 with \textit{MnlI} at -199, \textit{RsaI} at -116 and \textit{NdeI} at -97. (Numbering is with respect to the 5'→3' strand, and does not include overhanging 3' bases).

The vicl-8E fragment consisted of a pair of complementary oligonucleotides containing the sequence between -96 to -78 of \textit{Psa-v}. The oligos were made on an Applied Biosystems DNA synthesiser. Equimolar concentrations of oligonucleotide were annealed by heating at 70°C for 5 min (to disrupt secondary structure), then incubated at 60°C for 5 min and then allowed to cool slowly to room temperature. The annealed product was purified and eluted from a 20% (w/v) polyacrylamide gel (see below). The sequence of this 30 bp fragment is shown in Figure 3.3A.

Four DNA fragments were used as controls. These were:
- a 322 bp \textit{PvuII} fragment from pUC118 (not shown);
- a 192 bp \textit{HindIII/SalI} fragment from p4AREIGN (Olive \textit{et al.}, 1990; Olive \textit{et al.}, 1991) containing four copies of the GC-rich ARE (anaerobic responsive element);
- a 181 bp \textit{EcoRI/SalI} fragment from p4OCS-9b (kindly supplied by Jeff Ellis) containing four copies of the octopine synthase (OCS) enhancer element (Tokuhisa \textit{et al.}, 1990); and
- a 226 bp \textit{EcoRI/AvaI} fragment (Pharmacia, Australia) of the Epstein-Barr Virus genome, containing the binding sites of EBNA-1 (Epstein-Barr virus Nuclear Antigen-1) and CTF (CAAT-binding Transcription factor, Rawlins \textit{et al.}, 1985).

Essential regions of the 5'→3' strand of these sequences are shown in Figure 3.3.
Figure 3.3
Sequences of oligonucleotides and small DNA fragments used in gel-shift assays.

The DNA sequence of each promoter fragment is in upper case, added restriction sites are in lower case (sequences in the restriction site that are excluded by digestion are in brackets) and single stranded bases are in italics.

A. Fragment vicl-8E, contains the Psa-v sequence between -96 to -78. Both strands of DNA are shown and the cATGCATG motif is highlighted.

B. - D.
DNA sequences from p4AREIGN, p4OCS-9b and pEBV. Binding sites of GCBP-1 (Olive et al., 1991), OCSTF (Tokuhisa et al., 1990) and of EBNA-1 (Rawlins et al., 1985), respectively, are underlined. A double-headed arrow indicates the centre of the palindromic OCSTF recognition sites. The binding site of CTF is double-underlined, (Rawlins et al., 1985). Only the essential regions of the 5'→3' strand of these sequences are shown.

E. - F.
The sequences of the TAA and TAT fragments, respectively, containing gibberellic acid response elements. These sequences are derived from the α-amylase gene, amy1 (Gubler and Jacobsen, 1992).
<table>
<thead>
<tr>
<th>A.</th>
<th>BglI</th>
<th>5' Psa-v. vic I-8E</th>
<th>BamHI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>attagact</td>
<td>TATGCGATGGCACAAACATGG</td>
<td>atcc</td>
</tr>
<tr>
<td></td>
<td>tctag</td>
<td>ATACGTACCGTCTTGATCC</td>
<td>tagggt</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>B.</th>
<th>Hind III</th>
<th>SphI</th>
<th>4 X ARE</th>
<th>PstI</th>
<th>SalI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5'(a)agcttgcatgc</td>
<td>4X [CTGACGCCGCGCGCTCGGTCTTGCTTGCC]</td>
<td>ctgcaggtcga(c)3'</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>C.</th>
<th>Eco RI</th>
<th>SmaI</th>
<th>BamHI</th>
<th>4 X OCS</th>
<th>SalI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5'(g)aattccgtaggatccg</td>
<td>4X [TCGAGGGTCCACAAAAGCGTAAGCTGTTACGTACATGATG]</td>
<td>gtcga(c)3'</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>D.</th>
<th>Eco RI</th>
<th>EBV</th>
<th>Aval</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5'(g)aatcc</td>
<td>CTAA-70 bp-TGGATCTTGCTGTAGCATATGCTATCTCTAACGGA</td>
<td>tccccg 3'</td>
</tr>
<tr>
<td></td>
<td></td>
<td>TCCTCTAGATCCATATTGAGCTCTCAAGCCAAAAATTGATCT-60 bp-</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>E.</th>
<th>PstI</th>
<th>XbaI</th>
<th>5' α-amylase TAA</th>
<th>PstI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>gtctaga</td>
<td>CCGGCCGATAACAACTCCGG</td>
<td>tgca</td>
<td></td>
</tr>
<tr>
<td></td>
<td>acgtcagatct</td>
<td>GGGCCGCTATTGTTGAGGCC</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>F.</th>
<th>PstI</th>
<th>XbaI</th>
<th>5' α-amylase TAT</th>
<th>PstI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>gtctaga</td>
<td>CCGGACATATCCACTGGCCAAAG</td>
<td>tgca</td>
<td></td>
</tr>
<tr>
<td></td>
<td>acgtcagatct</td>
<td>CGGCTGTATAGGCTACCCGGGTTPC</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The TAA and TAT oligos (Figure 3.3), kindly provided by Frank Gubler and Jake Jacobsen, encompass control elements from the barley α-amylase gene. This gene is expressed in aleurone cells and is regulated by gibberellins and abscisic acid. (Gubler and Jacobsen, 1992).

All DNA fragments were isolated by electrophoresis in polyacrylamide gels. The concentration of the gels ranged from 3.5% - 20% (w/v) polyacrylamide (19:1 (w/w) acrylamide: bis-acrylamide), depending upon the fragment size; and were run in 1 X TBE buffer, (89 mM Tris buffer, 89 mM boric acid, 20 mM EDTA, pH 8.3, 5% [v/v] glycerol). The fragments were extracted from gel slices using two volumes of gel elution buffer (500 mM ammonium acetate, 10 mM magnesium acetate, 1 mM EDTA, 0.1% (w/v) SDS, pH 7.0), and incubating at 65°C for 10 min. Elution was continued at 37°C for 8 hours or 55°C for 1 hour, and the DNA was precipitated from the eluate by the addition of two volumes of ethanol.

3.2.4 Radioactive labelling of DNA fragments

Probes for gel-shift assays were prepared by labelling the ends of DNA fragments with DNA polymerase I (Klenow fragment), using [α-32P] dATP and/or [α-32P] dCTP. The labelled probes were then purified by electrophoresis in polyacrylamide gels, as described above, and diluted to 4000 cpm (Cerenkov)/µl (0.01 to 0.20 ng/µl) using sterile distilled water.

3.2.5 Binding of the nuclear protein extract to the radioactive probe

The protocol was based upon that described by Tokuhisa et al. (1990). Typically, 2.5 µg of nuclear protein were incubated with end-labelled fragments (4000 to 8000 cpm) for 20 min at 30°C, containing 2 µg poly (d(dC)), 5% (v/v) glycerol, 5 µg BSA, in a final volume of 20 µl binding buffer (10 mM Tris, pH 7.5, 50 mM NaCl, 1 mM EDTA, 1 mM DTT). After incubation, 1 µl of 0.1% (v/v) bromophenol blue was added to each reaction tube. The -773 to -423 probe, was initially incubated as above, but the proteins that bound to this fragment were much more abundant than those that bound to the other vicilin probes. Optimal resolution of binding to this fragment was achieved with 1 µg of nuclear protein and 4 µg of poly (d(dC)).
Competition assays were done as described above, except that the competitor DNA was incubated with protein extract for 10 min at room temperature prior to the addition of the DNA probes. Unless otherwise stated, the amount of competitor DNA used was a 100-fold molar excess of the probe DNA used. Proteins that were shown to bind to specific sites within the 5' region of Psa-v were given names that related to either their temporal regulation or the DNA content of their binding sites. Complexes that were due to non-specific DNA-protein interactions, or were not investigated were given names in lower case, referring to the fragment (a, b, c and d: pQ9, pQ8, pQ6 and pQ18, respectively) and the relative mobility (1,2,3 etc.) of the complex in the gel.

3.2.6 Polyacrylamide gel electrophoresis

Reaction mixtures were electrophoresed at 4°C for 2-4 hours on 4% (w/v) or 6% (w/v) polyacrylamide (40:1) gels, containing 5% (v/v) glycerol, in 0.25 x TBE buffer at 100V (approximately 20 mA), with buffer recirculation. Following electrophoresis, gels were dried and either exposed to X-ray film for 0.5 to 8 days, or to a phospho-imager screen (Johnston et al., 1990) for 4 to 24 hours.

3.2.7 RNA isolation and northern blotting

Pea cotyledons were harvested at different stages after flowering approximating 8-9 DAF (84 mg/pair), 11-12 DAF (127 and 194 mg/pair), 13-15 DAF (221 and 272 mg/pair) 16-17 DAF (327 and 369 mg/pair) and 21-22 DAF (576 mg/pair).

Total RNA was isolated following the procedure outlined by Verwoerd et al. (1989). In brief, the tissue was frozen with liquid nitrogen and ground to a fine powder with a mortar and pestle. Two volumes of NTES (100 mM NaCl, 10 mM Tris pH 8.0, 1mM EDTA, 1.0% SDS) and 1.5 volumes of PCI (phenol:chloroform:isoamylalcohol; 25:24:1 [v/v/v]) were added. The mixture was ground until well thawed; transferred to a test tube and then vortexed vigorously for 5 min. The mixture was centrifuged at 16 000 x g for 5 min. The upper, aqueous phase was transferred to a fresh tube, 0.1 volume of 2 M sodium acetate pH 5.8 and 2 volumes of ethanol, were added and the mixture was incubated overnight at -20°C. The mixture was centrifuged at 16 000 x g
for 10 min. The precipitated nucleic acid pellet was washed with 70% (v/v) ethanol, dried, and then dissolved in sterile water (1ml/g tissue). One volume of 4000 mM LiCl was added and the mixture was incubated on ice for 3 hours and then centrifuged at 16000 x g for 10 min to precipitate the RNA. The LiCl precipitation was repeated, the pellet washed with 70% (v/v) ethanol, dried and dissolved in sterile water (200 µl/g tissue).

RNA concentration was measured spectrophotometrically, assuming that one A260 unit represents 40 µg RNA/ml in a 1 cm path length. Total RNA (5 µg) was analysed using Northern blot analysis (Khandjian, 1986) with modifications as described by Higgins and Spencer (1991). The RNA was denatured in 50%(v/v) formamide and 17.5% (v/v) formaldehyde in the presence of 0.5 µg/µl ethidium bromide at 65°C for 5 min. After fractionation of the RNA on a 1.4% (w/v) agarose gel containing 5% freshly deionised formaldehyde and 1 X MOPS buffer (20 mM MOPS pH 7.0, 5 mM sodium acetate, 1 mM EDTA), the gel was photographed (to confirm even sample loading) and was blotted for 18 h onto a nylon membrane (Hybond-N, Amersham, Australia). After transfer, the damp membrane was irradiated with UV light for 2 min with a germicidal 15 W tube at a distance of 12 cm, to cross-link the RNA to the membrane.

A 740 bp \textit{Hpa I-Xba I} fragment containing +1110 to +1835 of the coding region of \textit{Psa-v} (Higgins \textit{et al.}, 1988; see Figure 2.3) was used to probe the blot. The fragment was labelled by oligo-labelling using a kit from Bresatec (Adelaide, SA) and denatured by incubation in 0.1N NaOH for 15 min at 37°C prior to use.

The membrane was pre-hybridised for 3 hours and hybridised overnight in the same solution (Khandjian, 1986; Khandjian, 1987) containing 1-5 x 10^5 cpm/ml of radioactively-labelled probe DNA. Hybridisations were carried out at 42°C. The membrane was washed with SSC + 0.1 % SDS at 68°C until the background radioactivity dropped to 100 - 200 counts per minute. The membrane was exposed to X-ray film at -80°C using a Hi-Plus (Du Pont) intensifying screen.
3.3 Results

3.3.1. Temporal regulation of DNA-binding proteins

The steady-state levels of vicilin mRNA in peas harvested at various stages of seed development were monitored by a northern blot (Figure 3.4). Vicilin messenger RNA was first (readily) detected in seeds 11 to 12 DAF, the steady state levels increased until 17 DAF and declined in 21 to 22 DAF seeds.

Four DNA fragments, containing the quantitative and seed-specific/temporal elements of the 5' region of Psa-v (Figure 3.1), encompassing the region from -773 to +1 of Psa-v, were used as probes in gel-shift assays to monitor the changes in DNA binding activity during early seed development. Nuclear protein extracts were used from seeds harvested at the same time as the northern blot experiment, ranging in developmental stages from 8 to 22 DAF. A variety of retarded complexes were formed with these fragments and that the relative abundance of these complexes varied during the development of the seeds.

Fragment -773 to -423 contains the major quantitative element in Psa-v (Chapter 2). At least 4 protein-DNA complexes, termed AT-2, a2, a3, a4, were formed between nuclear extract from pea seeds aged 11-12 DAF and this region of Psa-v (Figure 3.5). The formation of complex a2 was competed preferentially by high concentrations of poly dldC (Figure 3.5; 4, 6, and 8 µg), with a concomitant increase in the level of complex a4, suggesting that a2 may be a multimeric form of a4. The AT-2 and a3 complexes were not inhibited by lower concentrations of poly dldC, although they were reduced by the highest concentration.

DNA binding activity varied during seed development (Figure 3.6). Complex a4 was formed with extract from all seed stages, peaking around 12 DAF, whilst AT-2 and a3 binding activity showed greater variability; binding activity was not detected in the nuclear extracts from the youngest seed and also appeared lower at 16-17 compared to the adjacent stages of seed development. High poly (dldC) conditions were used, so that the a2 complex was not visible. Leaf nuclear proteins did not bind to this probe, suggesting that AT-2 and the other nuclear proteins may be specific to seeds.
Figure 3.4
Steady-state levels of vicilin mRNA in developing pea seeds.

RNA was isolated from developing seed at stages between 8 - 22 DAF. The RNA (5 µg) was fractionated on a formaldehyde-agarose gel, blotted onto nylon membrane and probed with a 32P-labelled 740 bp Hpa I-Xba I fragment of the coding region of Psa-v from plasmid pCW66.
Days after Flowering

<table>
<thead>
<tr>
<th></th>
<th>8-10</th>
<th>11-12</th>
<th>13-15</th>
<th>16-17</th>
<th>21-22</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Figure 3.5
Titration of DNA-protein binding to fragment -773/-425, containing the major quantitative element.

The top panel is a diagram of the DNA fragment -773/-425, that was used as the probe; the location of the AT-rich repeat sequences (black arrow heads and tails) are shown.

The bottom panel shows binding of nuclear proteins to the -773/-425 fragment. The probe was incubated in the absence of nuclear extract (first lane), or with 2.5 µg of nuclear extract from seeds 11-12 DAF. Reactions contained 2 to 8 µg of poly (dIdC), as indicated. DNA protein complexes AT-2, a2, a3, a4 and the free probe (fp) are indicated.
DNA probe
-773 to -425

Nuclear Extract
Poly dIdC

<table>
<thead>
<tr>
<th></th>
<th>-</th>
<th>+</th>
<th>+</th>
<th>+</th>
<th>+</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>2</td>
<td>4</td>
<td>6</td>
<td>8</td>
<td></td>
</tr>
</tbody>
</table>

AT-2
a2
a3
a4
fp
**Figure 3.6**  
Temporal regulation of DNA-protein binding to -773/-425.

Top panel: Diagram of the DNA fragment -773/-425, that was used as the probe; the location of the AT-rich repeat sequences (black arrow heads and tails) are shown.

Bottom panel: Binding of nuclear proteins to the -773/-425 fragment. The probe was incubated in the absence of nuclear extract (first lane), or with 1 µg of nuclear extract from seeds harvested at 8-10, 11-12, 13-15, 16-17 and 18-20 DAF (lanes 10, 12, 15, 17 and 20, respectively). The probe was also incubated with 2.5 µg leaf nuclear extract (last lane). All reactions contained of 4 µg of poly (dl-dC). DNA protein complexes AT-2, a3, a4 and the free probe (fp) are indicated.
The DNA sequence between -494 and -231 contains both the minor quantitative region (-424-327) and part of the seed-expression region from -306 to -251 (Figure 3.1). Stable-nuclear protein formed at least two complexes with the -494/251 fragment. These were termed AT-2 (which also bound to sequences in -79/426; see below) and B1 (Figure 3.2). Extended electrophoresis suggests that the B1 band may represent two or more complexes. The developmental pattern of AT-2 binding resembled that of AT-2 in -773 to -425 in the etiolated seed, but binding was detected in all extracts and peaked in the 10-12 DAF extract. A high-mobility complex, a3, migrating slightly slower than a1 (Figure 3.7) was also observed when leaf nuclear extracts were assayed.

Nuclear Extracts

Days after flowering

<table>
<thead>
<tr>
<th></th>
<th>10</th>
<th>12</th>
<th>15</th>
<th>17</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td>AT-2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>a3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>a4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>fp</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The remaining DNA fragment, containing nucleotide sequences between -773 to -425, also formed several DNA-protein complexes with seed nuclear extracts, none of which were observed with leaf nuclear extracts (Figure 3.9). The temporal pattern of one complex, termed TEMP/2, resembled that of TEMP-1. In contrast, a major complex, TEMP-3, was formed in nuclear extracts from all seed stages, even 8-10 DAF, peaking at 11-12 DAF. Although it appeared as if the TEMP-3 complex at 11-12 DAF might contain multiple complexes (Figure 3.6), extended electrophoresis was not able to resolve it into two distinct bands. A minor fast-migrating complex, a3, was formed with nuclear extracts from the younger seeds (6 to 15 DAF). Very faint complexes were also observed when leaf nuclear proteins were incubated with -86x1. These complexes migrated at rates slightly faster than TEMP-3 and slower than a3.
The DNA sequence between -424 and -231 contains both the minor quantitative region (-424/-307) and part of the seed-expression region from -306 to -231 (Figure 3.1). Seed nuclear proteins formed at least two complexes with the -424/-231 fragment. These were termed AT-2 (which also bound to sequences in -773/-425; see below) and b1 (Figure 3.7). Extended electrophoresis suggests that the b1 band may represent two or more complexes. The developmental pattern of AT-2 binding resembled that of AT-2 binding to the -773 to -423 probe; the binding proteins were observed in the extracts of older, but not the youngest, seeds. Binding of b1, like a4 binding, was detected in all extracts and peaked in the 11-12 DAF extract. A high mobility complex, L1, migrating slightly slower than b1 (Figure 3.7) was also observed when leaf nuclear extracts were assayed.

Another fragment, -231/-78, containing the remainder of the seed-expression region, also formed complexes with nuclear proteins. These, termed TEMP-1, c1 and c3, are shown in Figure 3.8. TEMP-1 binding appeared to be strongly developmentally-regulated: Binding activity was absent in 8-10 DAF extracts, and peaked in seeds between 11-15 DAF. When large quantities of nuclear extract were used, however, TEMP-1 binding could be also detected at low levels in older seed extracts (data not shown). Protein c2, like the other high mobility proteins a4 and b1, was observed in all except the 8-10 DAF seed extracts. The TEMP-1 and c2 complexes were not observed in leaf nuclear extracts, but a faster complex, c3, migrating almost with the free probe could be detected with all nuclear extracts tested, particularly the leaf extract (Figure 3.8).

The remaining DNA fragment, containing nucleotides from -96 to +1, also formed several DNA-protein complexes with seed nuclear extracts, none of which was observed with leaf nuclear extracts (Figure 3.9). The temporal pattern of one complex, termed TEMP-2, resembled that of TEMP-1. In contrast, a major complex, TEMP-3, was formed in nuclear extracts from all seed stages, even 8-10 DAF, peaking at 11-12 DAF. Although it appeared as if the TEMP-3 complex at 11-12 DAF might contain multiple complexes (Figure 3.9), extended electrophoresis was not able to resolve it into two distinct bands. A minor fast-migrating complex, d3, was formed with nuclear extracts from the younger seeds (8 to 15 DAF). Very faint complexes were also observed when leaf nuclear proteins were incubated with -96/+1. These complexes migrated at rates slightly faster than TEMP-3 and slower than d3.
Figure 3.7
Temporal regulation of DNA-protein binding to -424/-231.

Top panel: A diagram of the DNA fragment -424/-231, that was used as the probe; the location of an AT-rich repeat sequence (black arrow head) and a palindromic sequence (↔), as described in Figures 3.1 and 3.2, are shown.

Bottom panel: Binding of nuclear proteins to the -424/-231 fragment. The probe was incubated in the absence of nuclear extract (first lane), or with 2.5 µg of nuclear extract from seeds harvested at 8-10, 11-12, 13-15, 16-17 and 18-20 DAF (lanes 10, 12, 15, 17 and 20, respectively). The probe was also incubated with 2.5 µg leaf nuclear extract (last lane). All reactions contained of 2 µg of poly (dIdC). DNA protein complexes AT-2, L1, b1 and the free probe (fp) are indicated.
DNA probe
-424 to -231

Nuclear Extracts

Days after flowering

<table>
<thead>
<tr>
<th></th>
<th>10</th>
<th>12</th>
<th>15</th>
<th>17</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td>seed</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>leaf</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

AT-2

L1

b1

fp
Figure 3.8
Temporal regulation of DNA-protein binding to -230/-78.

Top panel: Diagram of the DNA fragment -230/-78, that was used as the probe; the location of the cATGCATG element (▲), vicilin box I, (containing a CACA element, ▼ and an octanucleotide box, ▽) and vicilin box II (▽) are shown, as described in Figures 3.1 and 3.2.

Bottom panel: Binding of nuclear proteins to the -230/-78 fragment. The probe was incubated in the absence of nuclear extract (first lane), or with 2.5 µg of nuclear extract from seeds harvested at 8-10, 11-12, 13-15, 16-17 and 18-20 DAF (lanes 10, 12, 15, 17 and 20, respectively). The probe was also incubated with 2.5 µg leaf nuclear extract (last lane). All reactions contained of 2 µg of poly (dIdC). DNA protein complexes TEMP-1, c2, c3 and the free probe (fp) are indicated.
DNA probe
-230 to -78

Nuclear Extracts

Days after flowering

TEMP-1

c2

c3

fp

--- seed --- leaf

10 12 15 17 20
Figure 3.9
Temporal regulation of DNA-protein binding to -96/+1.

Top panel: Diagram of the DNA fragment -96/+1, that was used as the probe; the location of the TATA box (T), CAAT box (C), cATGCATG element (▲) are shown, as described in Figures 3.1 and 3.2.

Bottom panel: Binding of nuclear proteins to the -96/+1 fragment. The probe was incubated in the absence of nuclear extract (first lane), or with 2.5 µg of nuclear extract from seeds harvested at 8-10, 11-12, 13-15, 16-17 and 18-20 DAF (lanes 10, 12, 15, 17 and 20, respectively). The probe was also incubated with 2.5 µg leaf nuclear extract (last lane). All reactions contained of 2 µg of poly (dIdC). DNA protein complexes TEMP-2, TEMP-3, d3 and the free probe (fp) are indicated.
DNA probe
-96 to +1

Nuclear Extracts

Days after flowering

| 10 | 12 | 15 | 17 | 20 |

TEMP-2

TEMP-3

d3

fp

The developmental stages of the mature embryos were not included because the steady-state levels of the mRNAs were not significantly different. The calculation, as done, demonstrated that TEMP-1 binding to the 34A78 fragment is reduced by the addition of a 100-fold molar excess of some other non-specific DNA, including fragments from pUC18 or the coding region of Ppa-v, indicating that the binding may be easily disrupted, or may have some affinity for other DNA sequences. Formation of the c1 and c2 complexes was not inhibited by the addition of unlabelled substrate DNA or non-specific DNA to the binding reactions (Figure 3.10), indicating that the binding was non-specific.
Incubation of nuclear extract obtained from seeds 12 DAF with proteinase K for 10 minutes at 37°C abolished binding to all four probes, whilst similar incubation with RNAse A had little or no effect, indicating that all the complexes were due to the formation of protein-DNA complexes (data not shown).

3.3.2 Location of binding sites in the seed-specific region and the minor quantitative element

CORRELATION OF DNA-BINDING ACTIVITY AND PSA-V GENE EXPRESSION

Since so many DNA-protein complexes were detected using nuclear extracts from seeds at different developmental stages, I decided to focus on the binding by proteins from seeds aged 11 to 15 DAF. The DNA sequences chosen for this more detailed study were from the seed-expression region, contained in the probes -420/-231 and -230/-78.

The developmental stages of the nuclear extracts were selected primarily because the steady-state level of vicilin mRNA in seeds 11-15 DAF increased most rapidly during this time (Figure 3.4), and so those seeds were ones in which transcription of Psa-v was likely to be at its highest levels.

BINDING IN THE SEED-EXPRESSION REGION

Competition assays were used to determine which of the DNA-protein complexes identified represented specific DNA-protein interactions. As shown in Figure 3.10, TEMP-1 binding to the -230/-78 Psa-v fragment was competed by a 100-fold excess of unlabelled substrate DNA, but not by a fragment of similar size, containing four copies of the Anaerobic Responsive Element (ARE) binding site (Olive et al., 1990; Olive et al., 1991), indicating that TEMP-1 recognises a specific sequence in this region. Additional experiments (data not shown) demonstrated that TEMP-1 binding to the -231/-78 fragment was reduced by the addition of a 100-fold molar excess of some other non-specific DNA, including fragments from pUC118 or the coding region of Psa-v, indicating that the binding may be easily disrupted, or may have some affinity for other DNA sequences. Formation of the c1 and c2 complexes was not inhibited by the addition of unlabelled substrate DNA or non-specific DNA to the binding reactions (Figure 3.10), indicating that the binding was nonspecific.
Figure 3.10
Sequence-specificity of TEMP-1 binding.

Top panel: Diagram of the DNA fragment -230/-78 that was used as the probe; the location of the cATGCATG element (▲), vicilin box I, (containing an octanucleotide box, □ and a CACA element, ▼) and vicilin box II (▼) are shown, as described in Figures 3.1 and 3.2.

Bottom panel: Binding of nuclear proteins to the -230/-78 fragment. The probe was incubated with nuclear extract from seeds 13-15 DAF, in the absence of competitor DNA (lane - ), or in the presence of a 100-fold molar excess of either unlabelled -230/-78 fragment (specific competitor), or the 4 x ARE fragment, containing four copies of the anaerobic responsive element from the maize Adh1 gene (non-specific competitor). DNA protein complexes TEMP-1, c2 and the free probe (fp) are indicated.
All of the conserved DNA motifs that are common to seed-expressed genes, and those specific to genes encoding 78 proteins, that are present in the Para v gene are located in the DNA sequence between -880 to -63 of Para v. Multiple conserved sequences include a CATGATG box (underlined) nucleotide in lower case), violin box (i and violin box1) (25 bp), which contains the octanucleotide box and the CAACA box. Sub-fractions of the entire region as well as oligonucleotides containing the CATGATG region of Para v and regulatory elements were used to determine the specific binding sites for the Para v competition experiments. The competition experiments indicated that the P200-116 fragment (Figure 3.11, lane 4), despite considerable identity between regions containing the CATGATG motif with binding to -200-78 (Figure 3.11, lane a), even when subbed into limited exon was used (data not shown). These data indicate that the original binding site(s) were located upstream of the CAACA and CATGATG motifs in the Para v promoter sequence.

To further define the TEMP-1 binding site(s), binding studies were done in which the binding regions used were much smaller fragments of the Para v promoter. The DNA fragment of -300-78, -999-116 and -119-77 each bound nuclear proteins that could be competed by the specific fragments. This indicates that TEMP-1 binding sites were located in -300-116 and -119-77.
All of the conserved DNA motifs that are common to seed-expressed genes, and those specific to genes encoding 7S proteins, that are present in the \textit{Psa-v} gene are located within the DNA sequence between -230 to -78 of \textit{Psa-v}. These conserved sequences include a cATGCATG box (mismatched nucleotide in lower case), vicilin box II and vicilin box I (33 bp), which contains the octanucleotide box and the CACA box. Sub-fragments of the entire region as well as oligonucleotides containing the cATGCATG region of \textit{Psa-v} and regulatory elements from a barley \(\alpha\)-amylase gene, were used in competition experiments to determine the location of the TEMP-1 binding site(s), and to determine whether TEMP-1 binds to regions containing the cATGCATG element (Figures 3.11).

In these experiments all competitors were present in 100-fold molar excess, compared to the -230/-78 substrate DNA fragment. The smallest fragment that completely competed TEMP-1 binding contained nucleotides from -230 to -116. The fragment -115/-97 also competed with binding, but to a lesser extent than the -230/-116 fragment. This suggests that TEMP-1 can bind to sequences in this region, albeit with lower affinity. This DNA fragment, -115/-97, contains part of vicilin box I, shown below, including the conserved CACA motif (underlined). Interestingly, a similar sequence, containing the \(\alpha\)-amylase gibberellic acid/abscisic acid response element TAACAAA (Gubler and Jacobsen, 1992) did not compete with TEMP-1 binding (Figure 3.11, lane f), despite considerable identity between the two sequences:

\[
\begin{array}{ccc}
-151 & 5' & \text{CCGGCGATAACAAAC-TCCGC} 3' -131 \\
-115 & 5' & \text{ACATTTTAACACACGTCCA} 3' -97 \\
\end{array}
\]

\textit{Psa-v}

\textit{amy1}

The region containing the CATGCATG element did not compete with binding to -230/-78 (Figure 3.11, lane e), even when 2000-fold molar excess was used (data not shown). These data indicate that the principal binding site(s) was (were) located upstream of the CACA and CATGCATG motifs in the \textit{Psa-v} promoter sequence.

To further define the TEMP-1 binding site(s), binding studies were done in which the binding regions used were much smaller fragments of the \textit{Psa-v} promoter. The DNA fragments -200/-78, -230/-116 and -115/-97 each bound nuclear proteins that could be competed by the specific fragment. This indicates that TEMP-1 binding sites were located in -200/-116 and -115/-97.
Figure 3.11
Localisation of TEMP-1 binding, by competition with fragments of -230/-78.

Top panel: Diagram of the probe -230/-78 and fragments of the sequence between -230 to -78 that were used as competitor DNA (a-e) and other DNA fragments (f-h) that were used as competitor. The location of the conserved elements in -230/-78 are as described in Figures 3.1 and 3.2. The gibberellic acid response elements TAACAA (▽) and TATCCAC (*) in the TAA and TAT sequences, respectively, are indicated. The GCBP-1 binding sites (●) in the fragment 4 x ARE are also indicated.

Bottom panel: Binding of nuclear proteins to the -230/-78 fragment. The probe was incubated in the absence of nuclear extract (first lane), or with nuclear extract from seeds 13-15 DAF, in the absence of competitor DNA (second lane) or in the presence of a 100-fold molar excess of unlabelled competitors from the *Psa-v 5*’ region: (a) -230/-78, (b) -230/-116, (c) -115/-78, (d) -115/-98, (e) -96/-78; or other sources (f) *amy1*, TAA, (g) *amy1*, TAT, (h) *Adh1*, 4 x ARE. DNA protein complexes TEMP-1, c2 and the free probe (fp) are indicated.
DNA probe
-230 to -78

Competitor DNA
a  -230 to -78
b  -230 to -116
c  -115 to -78
d  -115 to -96
e  -96 to -78
f  TAA
g  TAT
h  4x ARE

Competitor DNA Extract
- - a b c d e f g h
- + + + + + + + + + + +

TEMP-1
c2
fp

As expected, binding of TEMP-1 to PCNA at position -230 to -78 was competed effectively by the wild-type fragment -96 to -78. The absence of competitor DNA to locate the TEMP-1 binding region in the 20x ARE constructs shows that it is a specific sequence element.
The intensity of binding to these small fragments, however, was less than that for the full-length -230/-78 fragment, indicating that the length or composition of surrounding DNA sequences influenced binding.

Other DNA fragments; -230/-201 and -96/-78 (fragment vicI8E, with the cATGCATG motif) also bound nuclear proteins, but could not be competed by the Psa-v promoter fragment from -230 to -116, and so represented non-specific binding (data not shown).

Footprinting techniques, initiated to pinpoint the binding sites in the -231/-78, region, were unsuccessful, due to the instability and relatively low abundance of the TEMP-1 binding protein. Hence more detailed competition experiments were used to localise the binding sites. Fragments of the Psa-v promoter present in the plasmids pQ-σA to pQ-σH (Figure 3.2) containing deletions and/or base-pair substitutions of nucleotides in the region between -306 to +39 were used as competitor DNA to localise the TEMP-1 binding sites (Figure 3.12).

As expected, binding of TEMP-1 to nucleotides between -230 to -78 was competed effectively by the wild-type fragment -306/+36. Fragments containing substitutions within the region from -306 to -175, or within the region from -98 to -79 also competed for TEMP-1 binding to this region (Figure 3.12, constructs pQ-σA, pQ-σB, pQ-σC, pQ-σD and pQ-σH). In contrast, DNA fragments with mutations between positions -186 and -96 did not compete effectively (Figure 3.12, constructs pQ-σE, pQ-σF and pQ-σG). Thus, the sequences in the -186 to -98 region contain at least one TEMP-1 binding site. Sequence comparison of these three DNA sequences that were required for effective competition of binding shows that each one contains the consensus sequence 5' TTtttTtcAtTTTgacacat 3' (Figure 3.13), and so suggests that there may be three TEMP-1 binding sites in the region -186 to -98.

BINDING IN THE MINOR QUANTITATIVE ELEMENT.

Region -424 to -231 contains 76 bp of the seed-expression region as well as the minor quantitative element. Previous data showed that the region -227/-49 was not able to confer detectable expression on a chimeric CAT gene construct (Newbigin, 1988; and Appendix II). The region downstream of -306, however, was sufficient for detection of the expression of vicilin gene constructs. This suggested that a positive element might be located with -306 to -227.
Figure 3.12
Localisation of TEMP-1 binding, by competition with mutated *Psa-v* fragments.

Top panel: Diagram of the probe -230/-78 and DNA fragments that were used as competitors. The unlabelled competitor DNA fragments contained region -306 to +39 of *Psa-v* (wt) and mutated versions of this region (A to H) derived from constructs pQ20 and (pQ-) A-306/-272, B-306/-231, C-230/-194, D-198/-175, E-186/-150, F-153/-117, G-129/-96 and H-98/-79, respectively, containing nucleotide substitutions as indicated. The location of conserved sequences (as described in Figures 3.1 and 3.2) and the mutated sections (cross-hatched) are indicated.

Bottom panel: The percentage of TEMP-1 binding from 13-15 DAF seed extracts that was competed by a 10-fold molar excess of the unlabelled wild-type (wt) or mutated *Psa-v* fragments (A-H) with nucleotide substitutions located as described above. The level of binding to the probe was quantified using phospho-imagery (Johnston *et al.*, 1990). The relative effectiveness of each fragment (% competition of bound complex) is shown.
Figure 3.13
Comparison of sequences that contain regions required for TEMP-1 binding.

Four fragments from the 5' region of the *Psa-v* gene are aligned. Three of the sequences, -186/-159, -153/-117 and -129/-96, were required for TEMP-1 binding. The fourth, -115/-97, which is a sub fragment of -129/-96, was also recognised by TEMP-1, but with lower affinity. Numbers and letters in italics represent the position of the nucleotides immediately flanking the sequences that were mutated in constructs pQ-ΣE-186/-150, pQ-ΣF-153/-117 and pQ-ΣG-129/-96. Vertical lines represent identity between nucleotides in the fragment -153/-117 with nucleotides that were aligned in sequences -186/-150 (reversed) and -129/-97. Asterisks represent identity confined to aligned nucleotides in sequences -186/-150 and -129/-97. Nucleotides that were identical in all three aligned sequences are highlighted, in upper case. Nucleotides that were present in two of the three sequences are highlighted, in lower case. A consensus of the sequence present in all three DNA fragments is shown.

The relative affinity of TEMP-1 for each sequence, as defined by competition with subfragments (Figure 3.11) and internally mutated fragments (Figure 3.12) is indicated.
Comparison of sequences containing TEMP-1 binding regions.

<table>
<thead>
<tr>
<th>Psa-v sequence</th>
<th>TEMP-1 affinity</th>
</tr>
</thead>
<tbody>
<tr>
<td>-150 5'R gaagggtggtgagTTtgtTtgAaTTTgatgaattt</td>
<td>-186 ++</td>
</tr>
<tr>
<td>-153 5' caagtaatgTTttTtcAtTTTgccacTtcaattttg</td>
<td>-117 ++</td>
</tr>
<tr>
<td>-129 5' cacttcaaaTTttgTacAtTTTaacacacgtccat</td>
<td>-96 ++</td>
</tr>
<tr>
<td>-115 5' acAtTTTaacacacgtcc</td>
<td>-97 +</td>
</tr>
</tbody>
</table>

CONSENSUS TTttTTtcAtTTTgacacat
Figure 3.14
Titration of DNA-protein binding to fragment -424/-231, with specific and non-specific competitor.

Top panel: Diagram of the DNA fragment, -424/-231, that was used as the probe; it contains the minor quantitative element (424/-307) and part of the seed-expression region (-306/-231). The location of an AT-rich repeat sequence (black arrow head) and a palindromic sequence (↔) are shown.

Bottom panel: Competition of binding of nuclear proteins to the -424/-231 fragment. The probe was incubated in the absence of nuclear extract (first lane) or with nuclear extract from seeds 13-15 DAF. Unlabelled specific competitor DNA, -424/-231 was added in 20-fold, 50-fold and 100-fold molar excess, as indicated by the widening wedge. Non-specific competitor fragments from pUC118 were added in 20-fold, 50-fold and 100-fold molar excess, as indicated. Another non-specific competitor, containing four copies of the octopine elements (OCS), was added in 100-fold molar excess (last lane). No competitor DNA was added to the second and sixth lanes. DNA-protein complexes AT-2 and b1, and the free probe (fp) are indicated.
DNA probe
-424 to -231

Competitor DNA
-424 to -231 pUC118 OCS

Extract

AT-2
b1
fp

Competition experiments were done to identify end localized specific DNA binding sites in the -424/-231 probe and to determine whether the site(s) were located within the minor quantitative region, the possible quantitative region, or other sequences within the assay expression region.

Initial titration experiments showed that as little as 20-fold molar excess of -424/-231 competed with AT-2 binding to the -424/-231 probe, but up to 700-fold molar excess of the competitor DNA was needed to completely compete this lower molar excess with the probe (Figure 3A). The pUC118 did not reduce binding of AT-2, but it was less effective than the specific competitor. The pUC118 fragment had little or no effect upon b1 binding (Figure 3B). The kinetics of the inhibition of AT-2 binding by the pUC118 fragment suggested that there may be two populations present in the AT-2 band; one that is non-specific and was readily competed, and one that is specific and was not competed by the non-specific competitor, but competitively competed by different competitor sequences and extended upstream probe (data not shown).

Fragment -306/-231 did not compete AT-2 binding (lane c). The binding site was upstream, in the minor quantitative region (lane f). Furthermore, when pUC118 was used as a probe, an increased density from extracts could not be seen to above the maximum 4 times more than in control. This indicated that there was not an abundant nuclear protein that bound to the -424/-231 probe that would compete with AT-2. The -424/-231 probe was specific for the binding site for a non-specific binding site was contained (Figure 3A, lane c). This fragment contains the AT-rich repeats with the consensus GATAA/gATAAT. Although this fragment contained a portion of the pUC118 polylinker (XhoI and BamHI sites), it is unlikely that this contributed to binding, because the fragment -306/-231, which did not compete (lane f), also contains these polylinker sequences.
Competition experiments were done to identify and localise specific DNA binding sites in the -424/-231 probe and to determine whether the site(s) were located within the minor quantitative region, the putative quantitative region, or other sequences within the seed-expression region.

Initial titration experiments showed that as little as 20-fold molar excess of -424/-231 competed with AT-2 binding to the -424/-231 probe, but up to 100-fold molar excess of this specific competitor did not effectively compete the lower mobility band b1 (Figure 3.14). The presence of nonspecific pUC118 did reduce binding of AT-2, but it was less effective than the specific competitor. The pUC118 fragment had little or no effect upon b1 binding (Figure 3.14). The kinetics of the inhibition of AT-2 binding by the pUC118 fragment suggested that there may be two complexes present in the AT-2 band; one that is non-specific and was readily competed, and one that is specific and was not competed by non-specific competitor, but tests with different gel compositions and extended electrophoresis (data not shown) do not support this.

Fragment -306/-231 did not compete AT-2 binding, indicating that the binding site was located upstream, in the minor quantitative element (figure 3.15, lane f). Furthermore, when -306/-231 was used as a probe, nuclear proteins from seeds could not be readily detected binding to this fragment, even when maximum (4 times more than normal) amounts of nuclear extract were used. This indicates that this region does not contain a binding site recognised by abundant nuclear proteins from seed (Figure 3.16). The fragment did form the leaf complex L1 with approximately equal intensity to the binding to the -424/-231 probe. Although it has not been established whether this represents specific binding, these data suggest that the region -306/-231 may contain a binding site for a non-seed protein.

Competition by other sub-fragments of -424/-231 revealed that the AT-2 binding site was contained in the quantitative region, between -424 to -400 (Figure 3.15, lane c). This fragment contains two AT-rich repeats with the consensus GATAa/gATAAT. Although this fragment contained a portion of the pUC118 polylinker (Xba I and Bam HI sites), it is unlikely that this contributed to binding, because the fragment -306 to -231, which did not compete (lane f), also contains these polylinker sequences.
Figure 3.15
Localisation of AT-2 binding, by competition with fragments of -424/-231.

Top panel: Diagram of the probe -424/-231 and fragments of the sequence between -424/-231 that were used as competitor DNA (a-h) and a non-specific DNA competitor (i). The location of the conserved elements in -424/-231 are as described in Figures 3.1 and 3.2. The location of the binding sites (●) in the octopine synthase elements recognised by the factor OCSTF, in the fragment 4 x OCS are also indicated.

Bottom panel: The probe was incubated in the absence of nuclear extract (first lane), or with nuclear extract from seeds 13-15 DAF, in the absence of competitor (second lane), or in the presence of a 20 x M excess of unlabelled competitors from the Psa-v 5′ region: (a) -424/-231, (b) -424/-341, (c) -424/-400, (d) -399/-341, (e) -340/-309; (f) -308/-231, (g) -306/-261, (h) -260/-231 and from the OCS gene (i) the 4 x OCS fragment. DNA protein complexes AT-2 and b1 and the free probe (fp) are indicated.
DNA probe
-424 to -231

Competitor DNA

-424 to -231
-424 to -341
-424 to -400
-399 to -341
-340 to -309
-308 to -231
-306 to -261
-260 to -231
4x OCS element

Competitor DNA
- - a b c d e f g h i Extract
- + + + + + + + + + + +

AT-2
b1
fp
Figure 3.16
DNA-protein binding to -306/-231.

Top panel: Diagram of the DNA fragments -424/-231 and -306/-231, that were used as probes. The location of an AT-rich repeat sequence (black arrow head) and a palindromic sequence (↔), as described in Figures 3.1 and 3.2, are shown.

Bottom panel: Competition of binding of nuclear proteins to the fragments -424/-231 and -306/-231. The probes -424/-231 (first to fourth lanes) and -306/-231 (fifth to eighth lanes) were incubated in the absence of nuclear extract (first lane and fifth lanes) or with 10 µg of nuclear extract (4 times the usual quantity) from seeds harvested at 11-12 and 13-15, DAF (lanes labelled 12 and 15). The probes were also incubated with 10 µg of leaf nuclear extract (leaf). All reaction contained of 2 µg of poly (dIdC). DNA protein complexes AT-2, b1 ,L1 and the free probe (fp) are indicated.
When the small fragment, -424 to -400, containing the AT-2 binding site, was used as a probe, it was unable to form distinct complexes with AT-2, but a slightly larger fragment, -524 to -517, bound almost as effectively as the full-length probe, suggesting that additional length was required to stabilise the binding interactions (data not shown).

Competition experiments (Figure 3.17) showed that the fragment containing the binding site (-424 to -400, fragment b) from the minor quantitative element competitor was effective in competing with the major quantitative element competitor at -773/-423 of the AT-2 binding protein also specifically binds to a site in the major quantitative element (Figure 3.17, lane 5,9,10). The AT-rich binding sequence in the minor quantitative region of the target sequence present, but not in the major quantitative region.

The gel shows the different fragments of the repeats are highlighted, polymers are highlighted in blue and so on. Interestingly, the fragment, -517, did not compete by the specific binding sites. Since the AT-rich sequence between -424 to -570, in which only one specific DNA binding site was located, is repeated in the major regulatory element in -512 to -487 (depicted as an arrowhead in Figure 3.17), presumably the additional binding sequences are outside this region, (-773/-581), or are located in the remainder of the AT-rich repeats (-553/-513 and -456/-426, depicted as arrow tails).

It has been shown that AT-rich binding proteins can recognise dissimilar AT-rich binding sites, suggesting that recognition and binding is via the structure of the minor groove, rather than of specific sequences in the major groove (Solomon et al., 1986; and reviewed in Traverse, 1989, Churchill and Traverse,
When the small fragment, -424 to -400, containing the AT-2 binding site, was used as a probe, it was unable to form distinct complexes with AT-2, but a slightly larger fragment, -424 to -337, bound almost as effectively as the full length probe, suggesting that additional length was required to stabilise the binding interactions (data not shown).

Competition experiments (Figure 3.17) showed that the fragment containing the binding site (-424/-400, fragment b) from the minor quantitative element competed with the lowest mobility complex bound to sequences-773/-423 of the major quantitative element. A different fragment of similar size (-306/-231, fragment c) did not compete, indicating that the AT-2 binding protein also specifically binds to a site in the major quantitative element (Figure 3.17, lanes 5,9,10). The AT-rich binding sequence in the minor quantitative region of *Psa-v* is also present, with two mismatches, in the major quantitative region. The two regions on the fragments are compared below; the 10 bp repeats are highlighted, polylinker sequences are in lower case and single-stranded bases are italicised (bases -399, -398 were present on the complementary strand and so were not included in the numbering of the fragment -424/-400).

<table>
<thead>
<tr>
<th>adjacent regions</th>
<th>AT-2 binding sites in the 5' region of <em>Psa-v</em></th>
<th>location</th>
<th>element</th>
</tr>
</thead>
<tbody>
<tr>
<td>Xba Bam Sma</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ctgaggatcccccgACG GATAATAAT A GATAGATAAT TA</td>
<td>-424/-398</td>
<td>minor</td>
<td></td>
</tr>
<tr>
<td>TATATGAAAGTAaACG GATAATAAT A GATAAAATAA TA</td>
<td>-512/-486</td>
<td>major</td>
<td></td>
</tr>
</tbody>
</table>

Interestingly, the fragment -424/-400 competed with a2 binding but not a3 or a4, although the AT-2, a2, a3 and, to some extent, the a4, complexes were competed by the specific fragment -773/-425 (fragment d). This result indicates that the major quantitative element may contain other specific DNA-protein binding sites. Since the AT-rich sequence between -424 to -370, in which only one specific DNA binding site was located, is repeated in the major regulatory element in -512 to -457 (depicted as an arrowhead in Figure 3.17), presumably the additional binding sequences are outside of this region, (-773/-551), or are located in the remainder of the AT-rich repeats (-550/-513 and -456/-425, depicted as arrow tails).

It has been shown that AT-rich binding proteins can recognise dissimilar AT-rich binding sites, suggesting that recognition and binding is via the structure of the minor groove, rather than of specific sequences in the major groove (Solomon *et al.*, 1986; and reviewed in Travers, 1989; Churchill and Travers,
Figure 3.17
Location of AT-2 binding site in Psa-v-773/-237 and LegA1-668/-237 fragments.

Top panel: Diagram of the DNA fragments that were used as competitor DNA (a) LegA1-668/-237, (b) Psa-v-424/-400, (c) Psa-v-306/-231 and (d) Psa-v-773/-425. Fragments (a) LegA1-668/-237 and (c) Psa-v-306/-231 were also used as probes, as indicated. The approximate location of the AT-rich binding site in the LegA1 5' region is shown. The location of the AT-2 binding site in fragment (b), the palindromic region (↔) in (c) and other portions of the AT-rich repeats (black arrow heads and tails), as described in Figures 3.1 and 3.2, are also indicated.

Bottom panel: Competition of binding of AT-rich fragments from the vicilin gene Psa-v and the legumin gene LegA1. The LegA1 (first to fifth lanes) and Psa-v (sixth to tenth lanes) probes were incubated in the absence of nuclear extract (first lane and sixth lanes), or with 2.5 µg of nuclear extract from 13-15 DAF seeds. A forty-fold molar excess of each unlabelled competitor DNA fragment (a, b, c, d) was added, as indicated. No competitor DNA was added to the reactions in the second and seventh lanes. All incubations contained of 2 µg of poly (dl)C. DNA protein complexes UB, LB, AT-2, a2, a3 and a4, and the free probe (fp) are indicated.
### Competitor DNA

<table>
<thead>
<tr>
<th></th>
<th>DNA Probe</th>
<th>LegA1 -668 to -237</th>
<th>Psa-v -773 to -425</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>L: -668 to -237</td>
<td>- - a b c</td>
<td>- - d b c</td>
</tr>
<tr>
<td>b</td>
<td>P: -424 to -400</td>
<td>- + + + +</td>
<td>- + + + +</td>
</tr>
<tr>
<td>c</td>
<td>P: -306 to -231</td>
<td></td>
<td></td>
</tr>
<tr>
<td>d</td>
<td>P: -773 to -425</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### DNA Probe

<table>
<thead>
<tr>
<th>DNA Probe</th>
<th>Extract</th>
</tr>
</thead>
<tbody>
<tr>
<td>UB</td>
<td>+</td>
</tr>
<tr>
<td>LB</td>
<td>+</td>
</tr>
<tr>
<td>fp</td>
<td>+</td>
</tr>
</tbody>
</table>

### Image

- AT-2
- a2
- a3
- a4
- fp
Figure 3.18
Localisation of TEMP-2 and TEMP-3 binding.

A. The fragment -96/+1 was used as a probe and different DNA fragments containing parts of this region of Psa-v were used as competitors to locate the binding region of TEMP-2 and TEMP-3. The unlabelled competitor DNA fragments: -306 to +39 of Psa-v (wt), sub-fragments (3'k - 3'h, -96/+1) and mutated versions of this region (H, I, K) were derived from constructs pQ20, Δ3'k+13, Δ3'h-76, Δ3'i-88, Δ3'h-98, PQ18, Δ98-79, Δ78-73 and Δ61/-4, respectively. The sub-fragments contain Psa-v from position -306 down to position +13, -76, -88 and -98. The specific fragment (-96/+1) was also used as competitor. The fragments, H, I, K, contain -306 to +39 of the Psa-v gene, with nucleotide substitutions between sequences -98/-79, -88/-73 and -61/-4, respectively. The location of conserved sequences (as described in Figures 3.1 and 3.2) and the mutated sections (cross-hatched) are indicated.

B. Competition of TEMP-2 binding. TEMP-2 binding from 13-15 DAF seed extracts was competed by a 10-fold molar excess of the unlabelled wild-type (wt) fragment DNA, smaller fragments of the 5' region (-306/+13, -306/-76, -306/-88, -306/-88, -306/-98, -96/+1) or mutated Psa-v fragments (H, I, K, with nucleotide substitutions located as described above). The level of binding for each DNA-protein complex was quantified by phospho-imagery (Johnston et al., 1990). The relative effectiveness of each competitor fragment (% competition of bound complex) is shown.

C. Competition of TEMP-3 binding. TEMP-2 binding from 13-15 DAF seed extracts was competed by a 10 x M excess of the unlabelled wild-type (wt), smaller fragments of the 5' region (-306/+13, -306/-76, -306/-88, -306/-88, -306/-98, -96/+1) or mutated Psa-v fragments (H, I, K, with nucleotide substitutions located as described above). The level of binding for each DNA-protein complex was quantified by phospho-imagery (Johnston et al., 1990). The relative effectiveness of each competitor fragment (% competition of bound complex) is shown. The intensity of TEMP-3 binding was about 30 times greater than that of TEMP-2 binding (see Figure 3.9 for visual comparison).
A.

**Competitor DNA**

- **wt**
  - -306

- **3'k**
  - C

- **3'j**
  - T

- **3'i**
  - A

- **3'h**
  - G

- **-96/+1**
  - A

- **H**
  - C
  - T

- **I**
  - C
  - T

- **K**
  - C
  - T

**DNA probe**

B.

**% Competition of TEMP-2 Binding**

- **wt**
  - -306
- **3'k +13**
- **3'j -76**
- **3'i -86**
- **3'h -98**
- **-96 to +1**
- **H. -98 / 79**
- **I. -88 / 73**
- **K. -61 / 4**

**Competitor**

C.

**% Competition of TEMP-3 Binding**

- **wt**
  - -306
- **3'k +13**
- **3'j -76**
- **3'i -86**
- **3'h -98**
- **-96 to +1**
- **H. -98 / 79**
- **I. -88 / 73**
- **K. -61 / 4**

**Competitor**
An AT-rich binding region has been identified in a quantitative element of a pea legumin gene LegA (Shirsat et al., 1990; Meakin and Gatehouse, 1991), and a homologous gene, LegA1 (-549 to -316, Appendix II). Although this region does not contain the AT-2 binding consensus, or the smaller 10 bp repeat, it does contain 3 regions with considerable sequence identity (16/27 bp) to the AT-2 binding regions. So, competition experiments were done to determine whether the fragment containing the AT-2 binding region could compete with binding to the legumin element. Figure 3.17 (fourth lane) shows, however, that despite the homology between the AT-2 binding region and sites in the legumin binding region, that the fragment containing the AT-2 binding site (-424/-400) did not compete with binding to the LegA1 fragment. This suggests that a different nuclear protein bound to the AT-rich sequences in the legumin element. Alternatively, AT-2 might have bound the legumin element, and may have had a far greater affinity for the sequences in the legumin element and so was not competed by the -424/-398 fragment at the maximum (40-fold molar excess) concentration used in this experiment.

**BINDING IN REGION -96 TO +1**

The DNA sequence from -96 to +1 of Psa-v was not sufficient to generate detectable levels of vicilin in transgenic tobacco (Chapter 3). Nevertheless, this region may contain elements, in addition to the CAAT and TATA boxes, that are required for high-level or seed-specific expression in concert with other regulatory elements. Of the three complexes that bound in this region, both TEMP-2 and TEMP-3 appeared to bind specifically to regions within -96 to +1 (Figure 3.18). Binding by the high mobility protein d3 (Figure 3.9) was not competed by specific competitor, indicating non-specific DNA-protein interactions (data not shown).

A 100-fold molar excess of fragments -306/+39 and -306/+13 competed with the majority of the TEMP-3 binding to -96/+1, whilst the fragments -306/-76, -306/-88 and -306/-98 failed to compete (Figure 3.18, panel C). This indicates that the TEMP-3 binding site is located between -76/+13 and also that TEMP-3 had little or no affinity for sequences upstream of -76, including the TEMP-1 binding sites. Fragments -306/+39 with internal mutations between -98/-79 and -78/-73 (sH-/79/, sI-78/-73) also competed with TEMP-3 binding, indicating that the binding sites were not present in these regions. Mutation of -61/-4, however, prevented competition of binding, indicating that TEMP-3 binds in this region.
TEMP-2 binding, like that of TEMP-1, was readily competed by 100 M excess of non-specific competitor DNA. Competition with a lower concentration of competitor fragments (20-fold molar excess) revealed that, like TEMP-3, the binding site is localised in -75/+13, although at this concentration, binding was not fully competed by fragments containing -306/-76. In addition, TEMP-2 binding was not fully competed by the mutant -306/+39 fragment from ΣK-61/-4, indicating that TEMP-2, like TEMP-3 also binds in the region between -61 and -4 (Figure 3.18 B).

TEMP-2 and TEMP-3 bind in the same region, but differences between peak binding activities, at 13-15 and 11-12 OAF, respectively, make it less likely that these complexes may represent multimers of the same protein. Comparison of the activity of TEMP-2 with TEMP-1, however, suggests that they may represent the similar proteins. The sequence between -61/-4 contains a sequence, TTtttTatAtcTataaata, which resembles, to some extent the consensus TEMP-1 binding site (sites conforming to consensus in bold, TATA box underlined). Initial cross-competition assays that were done to determine whether TEMP-1 and TEMP-2 are the same protein were inconclusive.

3.3. Inhibition of DNA-binding proteins

The absence of observable AT-2, TEMP-1,-2 and -3 binding proteins in the leaf extract could be due to either absence of the proteins in the leaves or because of inhibition of the binding activity of these proteins by factors in the leaf extract. The binding activity of seed extract was observed after mixing with leaf extract to ascertain whether the presence of the leaf extract could inhibit DNA-binding activity. Figure 3.19 shows that both TEMP-1 and AT-2 binding were disrupted by the addition of leaf extract, whilst the non-specific binding complexes were not affected (TEMP-2 and -3 were not tested). The maintenance of the non-specific complexes makes it unlikely that this loss of binding activity was due to general proteolytic activity in the leaf extract. Also complexes L1, c3 and perhaps c2, were formed between the leaf nuclear extract and the probes -424/-231 and -230/-78, indicating that the leaf nuclear extract contains binding proteins.

To test whether the inhibition by leaf extract was due to a non-specific factor, the effect of leaf extract upon other DNA-binding proteins was assessed. The anaerobic responsive element binding protein, GCBP-1, (Olive et al., 1991) and the octopine synthase binding protein, OCSTF, (Tokuhisa et al., 1990) are
maize nuclear proteins that bind to target sites in the maize alcohol dehydrogenase gene Adh1 and the octopine synthase (OCS) gene from Agrobacterium tumefaciens, respectively. Epstein-Barr Nuclear Antigen (EBNA-1) is a mammalian protein that also binds to a specific target site, in the Epstein-Barr Virus genome (Rawlins et al., 1985).

The presence of leaf nuclear extract had little or no inhibitory effect upon the binding of EBNA-1 and OCSTF to their target sites (Figure 3.19). Binding of the ARE binding protein was also not inhibited by leaf extract (Mark Olive, pers. comm.). The protein LABF-1, which binds in AT-rich region of -666 to -237 of the pea legumin genes LegA1 and LegA (Appendix II; Meakin and Gatehouse, 1991), however, was inhibited by leaf extract (Figure 3.19). These results indicate that the inhibitory factor may specifically inhibit nuclear proteins binding to sites in these two seed storage protein genes.

3.4 Discussion

3.4.1 Organ specific and temporal regulation of DNA-protein binding.

DNA-protein binding of TEMP-1, TEMP-2 and TEMP-3 correlated with transcription of Psa-v

Several specific binding sites in the seed-expression region and the quantitative regions of Psa-v were identified. The proteins that bound to these sites were not detected in leaf extract, indicating that the binding of these proteins may be specific to seeds. In addition, the temporal regulation of these proteins appeared to be quite different.

AT-2 was detected during all except the youngest stage of seed development. TEMP-1 and TEMP-2 binding activity, on the other hand, was abundant in seed extracts from 11-12 and 13-15 DAF, but much reduced in older seeds. TEMP-3 binding activity was also highest in extracts from 11-12 DAF seeds, although it was observed at lower levels in older tissue and even the 8-10 DAF extract.

The peak TEMP-1, TEMP-2 and TEMP-3 binding activities correlate with the initial accumulation of vicilin mRNA in developing seeds between 11 to 15 DAF (Chandler et al., 1984; Figure 3.4). The stages at which the steady-state levels of a mRNA class accumulate, however, do not pinpoint the developmental
Figure 3.19
Inhibition of binding proteins from seeds by pea leaf extract.

Upper panels, from left to right: Binding of nuclear proteins to the Psa-v fragments -773/-425, -424/-231 and -230/-78. Each panel shows binding by 2 µg of nuclear protein from 13-15 DAF seed extract (left), leaf and seed extract (2 µg of each extract, centre) and 2 µg of leaf extract only (right). The complexes indicated by triangles for each panel are AT-2, AT-2 and TEMP-1, respectively.

Lower panels, from left to right: Binding of nuclear proteins to the EBNA element of the Epstein Bar virus (Eco RI -Avai I fragment), the octopine synthase element of OCS (4 x OCS probe) and the positive element, -668/-237, of the pea legumin gene, LegA1. The EBNA element was incubated with 1 unit of EBNA-1 extract, with or without 2 µg of pea leaf extract, as indicated. The upper band represents specific binding by EBNA-1, the lower band is free probe. The OCS element was incubated with 5 µg of maize extract with and without 5 µg of pea leaf extract. The complex formed contains OCSTF bound to both OCS sites (complex lb in Tokuhisa et al., 1990), the free probe is not visible. The LegA1 probe, was incubated with 2 µg of 13-15 DAF seed extract with and without 2 µg of leaf extract, as indicated. The top complex represents specific binding (UB in Appendix II). The specific DNA-protein complexes are indicated by triangles.

All reactions contained 5 µg of bovine serum albumin. The reactions with probe -773/-425 had 4 µg of poly (dIdC), all other reactions contained 2 µg of poly (dIdC).
<table>
<thead>
<tr>
<th>Psa-v probes</th>
<th>-773 to -425</th>
<th>-424 to -231</th>
<th>-230 to -78</th>
</tr>
</thead>
<tbody>
<tr>
<td>leaf extract</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>seed extract</td>
<td>+</td>
<td>+</td>
<td>-</td>
</tr>
</tbody>
</table>

**DNA probes**

<table>
<thead>
<tr>
<th>DNA probes</th>
<th>EBNA element</th>
<th>OCS element</th>
<th>legumin element</th>
</tr>
</thead>
<tbody>
<tr>
<td>leaf extract</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>EBNA-1 extract</td>
<td>+</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>maize extract</td>
<td>-</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>seed extract</td>
<td>-</td>
<td>-</td>
<td>+</td>
</tr>
</tbody>
</table>

Temporal regulation of gene expression is achieved by the binding of nuclear proteins to specific DNA sequences. The EBNA and OCS elements are involved in this process. Temporal regulation of gene expression is achieved by the binding of nuclear proteins to specific DNA sequences. The EBNA and OCS elements are involved in this process.
stages of when transcription is at a peak, because of the relationship between mRNA transcription and degradation. In pea, three related classes of vicilin mRNAs have been identified from a study of cDNA clones (Chandler et al., 1984). The vicilin gene *Psa-v* is related to the gene class represented by the clone pPS15-84. *In vitro* transcription assays, analysing RNA synthesis of isolated pea nuclei in reaction media containing radioactive nucleotides, indicated that the transcription rate of mRNA corresponding to this vicilin class peaked at 11 DAF (Beach et al., 1985). The correlation between the high binding activity of the TEMP proteins during the early stages of expression of *Psa-v* supports the hypothesis that they may have a role in the positive regulation of the transcription of this vicilin gene.

Temporal regulation has been observed for other seed-specific DNA-binding proteins that bind to specific sites in the 5′ regions of seed storage protein genes. Binding of LAFB1 to the pea *LegA* legumin gene between -549 to -319, resembles that of TEMP-1 and TEMP-2, in that it was absent in seed nuclei prior to transcription of the legumin gene, but was detected in abundance during early stages of gene transcription, declining before the peak in *LegA* mRNA levels (Meakin and Gatehouse, 1991). Similarly, run-off transcription assays with isolated nuclei showed that the relative rate of transcription of the soybean *Le1* lectin gene correlated well with the binding activity of the 60 kDa DNA-binding protein to AT-rich sites in the gene (Jofuku et al., 1987).

There are nuclear proteins which bind to specific sequences in seed storage protein genes but which do not correlate with the onset of mRNA accumulation. The activities of soybean embryogenesis factors SEF2 and SEF4, which bind in the (7S) α,β-conglycinin gene, differ from AT-2, TEMP-1,-2 or -3. The binding activities of both SEF2 and SEF4 were only observed in seeds that were already transcribing of α,β-conglycinin mRNA. Binding of SEF3 was weakest in nuclear extracts prior to mRNA synthesis and strongest when mRNA levels (and not necessarily transcription rates) were greatest (Allen et al., 1989; Lessard et al., 1991).

Some of these differences in correlation between gene activation and DNA binding activity may reflect different roles in transactivation of gene expression during seed maturation. That is, nuclear proteins that are active at the early stages of gene expression may be required for activation of transcription, whilst different proteins that are active at later stages of seed development might interact with the initial proteins and enhance gene expression, perhaps at
specific temporal stages. Since storage protein genes are basically switched on at the same time as general seed metabolism increases, one might also expect that several other, more general transcription factors, would be switched on, and that these might contribute to the expression of various seed-specific genes.

Given the strong temporal regulation of TEMP-1, -2 and -3, and the more "constitutive" activity of AT-2, I propose a model for the regulation of \( Psa-v \) expression. In this model, binding of one or more of the TEMP-1, TEMP-2 and TEMP-3 proteins is essential for activation of \( Psa-v \) transcription. The rate of transcription is enhanced by the (previous) binding of AT-2 (and perhaps the other specific binding proteins, a2 and a3) to upstream regions of the \( Psa-v \) promoter. This enhancement could be a result of stabilising interactions between AT-2 and proteins of the transcription initiation complex. Because, in this model, transcription requires the presence of the TEMP proteins, the decrease in the binding of TEMP-1, -2 and -3 to the \( Psa-v \) promoter during seed development would lead to the reduction, or prevention, of the enhancement of gene expression by AT-2, and a subsequent reduction (as is observed) in transcription of \( Psa-v \) in more mature seeds.

**INHIBITORS OF PROTEIN-BINDING TO DNA**

DNA-binding inhibitors have not been reported for other seed-specific binding systems though they have been observed in both mammalian (see Chapter 1 for review) and plant transcription systems in other organs. The carrot protein EGBF-1, for example, is present in untreated root tissue, and specifically binds to an AT rich site in an extensin gene (Holdsworth and Laties, 1989b). Following wounding, the activity of EGBF-1 declines to undetectable levels. Mixing of nuclear extracts from wounded and unwounded roots demonstrated the existence of a specific wound-induced inhibitor. Titration and reconstitution experiments found that the EGBF-1 protein was not present in wounded root extracts (Holdsworth and Laties, 1989a), indicating that EGBF-1 activity is modulated by the regulation of both the DNA-binding activity and the level of abundance of the protein.

Binding activity of certain nuclear proteins to their specific target sites in the seed storage proteins genes \( Psa-v \) and \( LegA1 \) may be repressed in non-seed tissue. Leaf nuclear extract was found to contain a factor that prevented binding of TEMP-1 and AT-2 to their respective target sites, as well as protein-binding to an AT-rich fragment from \( LegA1 \) (Figure 3.19).
The loss of binding of AT-2 and TEMP-1 might have been caused by general proteolytic activity in the leaf extract. The leaf extract, however, had little or no effect on the binding of two maize nuclear proteins or a mammalian protein to their respective binding sites, suggesting that the inhibitory factor specifically modified or bound to the nuclear protein from seeds (Fig 3.19). Hence, the lack of binding activity, in leaves and perhaps, also, the changes in activity in seeds at different stages of development, may be controlled by regulating activity of the DNA-binding proteins. This form of regulation might be in addition to, or instead of, the transcriptional regulation that controls the abundance of these DNA-binding proteins.

Not all leaf nuclear extracts contain DNA-binding inhibitors. In similar experiments with a maize zein gene, the seed-specific beta-1 protein, which binds between -67 to -46 of the beta-zein gene, was not present in seedling tissue, but the seedling extract did not inhibit binding activity. These results indicated that the seedling tissue contained undetectable levels of beta-1, rather than containing an inhibitor that prevented the formation of the beta-1/DNA complex (So and Larkin, 1991).

In general, the inhibition of transcriptional activation, by the prevention of binding of a nuclear protein to its recognition site, can involve mechanisms that mask or inactivate the active site by either modification of the DNA-binding protein, or the binding of another protein (see Figures 1.2 and 1.3). For example, the DNA-binding activity of many nuclear proteins that bind to elements in light regulated genes are affected by phosphorylation status. The binding activities of most of these proteins appear to be inactivated by de-phosphorylation. These include SBF-1 (Harrison et al., 1991), GBF-1 (Klimczak et al., 1992) and 3AF3 (Sarokin and Chua, 1992). AT-1 binding, however, is inactivated by phosphorylation and reactivated by de-phosphorylation (Datta and Cashmore, 1989). A protein kinase of the type (casein kinase type II) that participates in controlling the binding of AT-1, has recently been purified and characterised from nuclei from pea plumules (Li and Roux, 1992). The binding activities of other eukaryotic DNA-binding proteins, such as cMyb, cJun, Max and SRF, are also affected by phosphorylation (reviewed in Hunter and Karin, 1992).
In addition to modification by phosphorylation, DNA-binding activity of a protein can be inhibited by the binding of second protein. For example, binding of the factor MyoD1 is required for the transactivation of muscle-specific genes (Davis et al., 1987). Expression of these genes is forestalled in non-muscle tissues by a high concentration of an inhibitory protein, Id, which binds to MyoD1 and prevents it from binding to target DNA sites in the promoters of the muscle-specific genes. Muscle gene activation is initiated by a fall in the concentration of Id, rather than an increase in the concentration of MyoD1 (Benezra et al., 1990). The DNA-binding of the cAMP response binding protein CREB (Foulkes et al., 1991; Foulkes et al., 1992a), the proto-oncogenic cJun (Nakabeppu and Nathans, 1991), and the liver-enriched transcriptional activator LIP (Descombes and Schibler 1991 (Descombes and Schibler, 1991), are also inhibited by the interaction and binding of specific inhibitory proteins (reviewed by Foulkes and Sassone-Corsi, 1992b).

Thus, the inhibitor of TEMP-1, AT-2 and LABF-1 may be a protein kinase or a phosphatase, similar to those that regulate phosphorylation of light-gene binding proteins in pea. Alternatively, it might be another protein/compound that inhibits binding of these seed-specific proteins.

3.4.2 Comparison of DNA-binding sites in the seed-specific region of 
Psa-v with those in other 7S genes.

TEMP-1 MAY BIND TO THREE SITES

Three regions that were required for the binding of TEMP-1 were identified in the seed-expression region of Psa-v, located between positions -186/-150, -153/-117 and -129/-96 (Figure 3.12). Because the three fragments that did not fully inhibit binding contained adjacent or overlapping mutations, the data does not clearly distinguish whether there are two or three (or more) individual TEMP-1 binding sites, or one large binding site between -186/-150 (\(\Sigma\ E\ -186/-150\)) and -153/-96 (\(\Sigma\ F\ -153/-117, \Sigma\ G\ -129/-96\)). Since both the DNA fragments -230/-116 and -115/-97 were able to bind TEMP-1, albeit weakly (data not shown), this suggests that both fragments contain a TEMP-1 binding site.
Constructs ΧF-153/-117 and ΧG -129/-96 share a common mutated region, between -129/-117. It is possible, therefore, that ΧG -129/-96 contains two mutated binding sites, between -129/-117 and -115/-96. Alternatively, the sequence between -153/-117 (mutated in ΧF-153/-117) may contain a TEMP-1 binding site.

Sequence comparison of the three regions required for TEMP-1 binding supports the hypothesis that each sequence may contain an individual TEMP-1 binding site, rather than part of one large binding site. Each fragment contains the consensus sequence TTttTcAtTTTgacat (Figure 3.13, conserved bases in upper case). The smaller fragment, -115/-97, contains most, but not all, of this sequence, which could explain why TEMP-1 had lower affinity for this sequence.

In addition to binding sites in the seed-expression region, the remainder of the seed-specific region, sequences between -96/+1, was shown to contain TEMP-2 and TEM-3 binding sites. These were located between positions -61 and -4 (Figure 3.18).

DNA-binding sites in Psa-v and two other genes encoding 7S proteins share little similarity

DNA sequences that bind nuclear proteins have been identified in regions containing seed-specific and temporal regulatory elements of other genes encoding 7S proteins; phaseolin (Allen et al., 1989) and α',β-conglycinin (Kawagoe and Murai, 1992), as well as vicilin. The presence of both different and multiple binding sites in these regions suggests that they may contain many cis-acting elements. The 5' regions of these three genes all share sequence identity, especially around the conserved vicilin box I sequence (Newbiggin, 1988; Chee and Slightom, 1991). In addition, these related genes are all regulated in a similar manner, both temporally and spatially, which suggests that they may be regulated by similar cis- and trans- acting factors. Despite these similarities, on first appraisal there appears to be little sequence identity between the observed DNA-binding sites in these genes. Most sites in one gene are not bound by nuclear proteins in vitro or not present in the other genes, however they all share T-rich binding regions.
The seed-specific region of the α′,β-conglycinin gene, *Gma-α′* (Chen *et al.*, 1986; Chen *et al.*, 1988) contains SEF3, and SEF4 binding sites, with core recognition motifs AACCCA and a/gTTTTTa/g, respectively (Allen *et al.*, 1989). A region of the phaseolin gene *Pvu-β* that is similar in function (Bustos *et al.*, 1991; Burow *et al.*, 1992) and DNA sequence (50-57% identity, Chee and Slightom, 1991), is located between positions -295 and +1. This region contains binding sites for the CAN, CA-1, AG-1, C1 and TATA-binding proteins (Kawagoe and Murai, 1992). The seed-specific region of *Psa-v*, as discussed above, contains binding sites for the TEMP-1, -2 and -3 binding proteins.

Comparison of the DNA sequences and the binding sites between these three seed-specific regions shows that *Psa-v* does not contain a SEF3 binding sequence (Higgins *et al.*, 1988), and, while *Pvu-β* does contain an AACCCA sequence at position -209, nuclear proteins have not been observed binding to that region (Kawagoe and Murai, 1992). At least two AACCCA sites are required for SEF3 binding to occur *in vitro* (Allen *et al.*, 1989) and so the single SEF3 site in *Pvu-β* was probably insufficient. Alignment of the 5′ regions of these genes (Chee and Slightom, 1991), however, shows that a TEMP-1 binding site from *Psa-v*, and two CA-1 binding sites from *Pvu-β* are located in very similar positions to the SEF3 binding sites, in relation to other conserved sequences and to the transcription start site. Despite this, comparison of these sequences suggests that TEMP-1 may recognise different sites to SEF4 and CA-1 (see inset below: TEMP-1, CA-1 and SEF3 binding sites are highlighted, the AG-1 site is in italics, while vicilin box II in *Psa-v* is underlined).

![Comparison of binding sites in three genes encoding 7S genes](image)

Indeed, the TEMP-1 binding consensus more closely resembles the T-rich binding sites that are bound by SEF4 (a/gTTTTTa/g) and AG-1 (TTGc/tCTTTTT). Coincidentally, three copies of each of these three sequences have been detected in the 5′ region of their respective genes. Although a TEMP-1 consensus site in *Psa-v* is located at the same position as an AG-1 site in *Pvu-β* (between the two CA-1 sites, see above), on the whole
the binding sites are located in different regulatory regions of these 7S genes, suggesting that these protein binding sites may have different functions. For instance, all the TEMP-1 consensus sequences are present in the seed-expression region of *Psa-v* (between -186 to -96), while the other T-rich binding sites in *Pvu*-β and *Gma*-α' are more widely dispersed. The SEF4 sites are at -690, -340 and -115 (Allen *et al.*, 1989), upstream of, and in minor and major quantitative elements, respectively (Chen *et al.*, 1986), and two of the AG-1 sites are located at -376 and -356 in an element that reduces expression in seeds, whilst the third is present at -191, in a negative element that repressed expression in stems and roots (Burow *et al.*, 1992; Kawagoe and Murai, 1992).

Nuclear protein CAN-1 bound to three CAnnTG sites in the temporal and tissue-specific regions of the phaseolin gene *Pvu*-β (Kawagoe and Murai, 1992). Despite the presence of homologous sequences, CAacTG, CAaaTG and CAtaTG, in the 5′ regions of *Psa-v* (at -98 and -76) and *Gma*-α' (at - 65) these were not observed to be bound by specific nuclear proteins (Figures 3.11; 3.12; Allen *et al.*, 1989; Lessard *et al.*, 1991).

The disparity of binding between these related genes poses a question about the roles, if any, these DNA-binding proteins have in transcriptional regulation. Some of the DNA-binding proteins may have no discernible role, and so may act as proteinaceous red-herrings when searching for trans-acting factors. The proteins SEF3, and perhaps SEF4, fall into this class (Fujiwara and Beachy, 1992). One could suggest that other binding sites and the DNA-binding proteins, ones that do contribute to the regulation 7S gene expression in these three 7S legume genes, have diverged significantly. This seems unlikely, however, since the 5′ regions of all these genes confer seed-specific expression on reporter genes in transformed tobacco (Newbigin, 1988; Appendix II; Bustos *et al.*, 1991; Fujiwara and Beachy, 1992; Chamberland *et al.*, 1992), which demonstrates that the cis-acting elements in these genes can be recognised by the trans-acting factors from widely diverged species.

How then does one explain the lack of consensus binding sites within such similar genes? It is possible that the consensus sequences recognised by these proteins are fairly loose, or that the proteins bind with relatively low specificity, so that some of these different DNA sequences might be bound by homologous proteins. This hypothesis could be tested by analysing the binding affinity of the nuclear proteins from seed extracts of the different species to the defined binding sites. Alternatively, since all of these DNA-
binding proteins that have been examined appear to be regulated temporally to some degree (Lessard et al., 1991; Figures 3.6 to 3.9), it is possible that some of these observed differences in binding are due to developmental differences between the nuclear extracts, and hence represent binding to and by different temporal factors, or changes in sequence affinity, mediated by protein modification.

TEMP-1 BOUND TO CONSERVED SEQUENCES IN PsA-v.

Curiously, with a few exceptions, specific binding of nuclear proteins has not previously been observed in sequences that are conserved in the 5' regions of legume seed storage protein genes; these include the CATGCATG motifs, the CATGCATG-containing legumin box, and vicilin box I (Jofuku et al., 1987; Allen et al., 1989; Bustos et al., 1989; Jordano et al., 1989; Riggs et al., 1989; Lessard et al., 1991; Meakin and Gatehouse, 1991; Fujiwara and Beachy, 1992; Kawagoe and Murai, 1992; Appendix II). A multimerized CATGCATG oligo also failed to bind nuclear proteins in vitro (Riggs et al., 1989). This is contrary to the situation observed for highly conserved elements in other genes such as light-regulated (Gilmartin et al., 1990; and Chapter 1, for review) and heat shock genes (reviewed in Gurley and Key, 1991).

In this study, however, two regions required for TEMP-1 binding, located between -153/-117 and -129/-97, encompass all of (the shorter, 33 bp) vicilin box I of PsA-v (Figures 3.12, 3.13). Comparison of the three TEMP-1 binding regions, however, suggests that it is the adjacent T-rich sequences, rather than the conserved octanucleotide box and the CACA box, that may contain the core recognition sequences. Nevertheless, the short region between -115/-97, which lacks the T-rich region but has the CACA motif, was able to compete TEMP-1 binding, albeit less effectively than the fragment -230/-116, which contained the remaining TEMP-1 sites (figure 3.8a). In addition, when this fragment was used as a probe, weak, but specific binding of nuclear proteins could be detected (not shown), indicating that TEMP-1 could recognise and bind to this sequence.

One of the two CACA motifs in Pvu-β is also present in a CA-1 binding sites, but binding was not detected to an identical CACA motif in the vicilin box sequence (Kawagoe and Murai, 1992). A 22 bp sequence from a zein promoter, that has the potential to form a stem loop, and contains two CACA-like sequences, is also bound by a seed-specific protein (Maier et al., 1987). These results suggest that DNA-proteins may recognise CACA sites, including
those with vicilin box I, but that additional bases are required for effective binding.

SEF4 binds to three sites within the Gma-α' gene. One of these sites, -115/-105 (Allen et al., 1989), is located in vicilin box I of the α',β-conglycinin promoter. This sequence is almost identical (9 bp/11 bp) to the sequence between -116/-106 of Psα-v, which is in the fragment -129/-97 which binds TEMP-1 (Figure 3.9b). Mutational analysis indicates that this SEF4 site is not essential for gene expression (Fujiwara and Beachy, 1992). These results do not conclusively show, however, that SEF4 does not have some transcriptional role because of the gene constructs that were analysed, either the both upstream SEF4 sites were left intact, or regulatory elements from the CaMV 35S were included, and so these sites may have compensated for the loss of the proximal SEF4 site.

The third TEMP-1 binding site is located in 20 bp of the 21 bp sequence of vicilin box II. This sequence is also present in other genes encoding 7S genes including the pea convicilin genes Cvc-a and Cvc-b, soybean Gma-α' and Vicia faba Vic-1 although it is not present in Pvu-β (Newbigin, 1988; and Appendix I). Another SEF4 binding site, located between -325/-340 (Allen et al., 1989), is contained in the 3' half of vicilin box II in the Gma-α' gene, indicating that vicilin box II may contain a conserved DNA-protein binding site. The nuclear protein CAN-1 has been observed binding in the extended vicilin box I sequence of Pvu-β, but, as noted above, specific protein binding was not observed in sequences which conformed to the binding consensus sequence, CAnnTG, in the 5' regions of the two other 7S genes, suggesting that this binding protein may not be present in pea.

The TEMP-2 and -3 binding sites are located between -61 and -4, and are downstream of the conserved vicilin II, I and CAAT box sequences. Although either of these may represent TATA binding proteins, like the TATA-binding protein that binds between -43/-27 of Pvu-β (Kawagoe and Murai, 1992), the strong temporal regulation of these proteins, correlating closely with transcriptional regulation of Psa-v, suggests that one or both may have a role in regulating expression during seed development.
SPECIFIC NUCLEAR PROTEINS DID NOT BIND TO THE CATGCATG ELEMENT

Despite the fact that the CATGCATG motifs are required for both high-level and seed-specific expression (CATGCATG in LeB4, Baumlein et al., 1992; CATGCATG and CATGCACa in Gma-α, Chamberland et al., 1992; Fujiwara and Beachy, 1992), nuclear proteins have not been shown to bind to these sites in legume seed storage protein genes (Allen et al., 1989; Riggs et al., 1989; Lessard et al., 1991; Meakin and Gatehouse, 1991; Kawagoe and Murai, 1992). DNA-binding was noted to the sequence -96 to -78 of Psa-v, which contains a CATGCATG sequence (tATGCATG) however competition assays found that this binding was non-specific (Figure 3.11 and data not shown).

Nuclear proteins have been shown to bind specifically to an oligonucleotide containing two CATGCA sequences, representing sequences between -53 to -46 of a seed storage protein gene encoding napin, from Brassica napus). It was not established, however, whether the CATGCA sites were critical for binding (Gustavsson et al., 1991). The nuclear proteins that bound to this sequence, or those of similar mobility, were also extractable from both seeds, prior to napin expression, and leaves (Gustavsson et al., 1991). This suggests that it is unlikely that such factors have a major role in regulating seed-specific expression. It is possible, however, that the transactivation function of the putative CATGCA binding proteins could be selectively inhibited in cells prior to napin expression, without affecting DNA-binding activity.

Alternating purine/pyrimidine base pairs have the potential to form a Z-DNA conformation, and such sequences have been implicated with influencing gene expression (An, 1987; Nathan et al., 1988). Hence it has been proposed that CATGCATG motifs may function by adopting an altered configuration that enhances the recognition or interaction of other regulatory factors (Riggs et al., 1989; Aryan et al., 1991). Alternatively, CATGCATG binding proteins, if they exist, may be labile or may not be required at high concentrations to activate gene expression and hence would not be readily detected in vitro, using crude nuclear extract.
3.4.3 AT-rich binding sites are prevalent in quantitative elements of seed-storage gene promoters

AT-2 binding sites were located in each quantitative element.

AT-2 bound to two AT-rich sequences in the 5' region of Psa-v. Each sequence was located in a quantitative element; at positions -512/-486 and -424/-398. AT-rich binding sites have been located in many seed storage protein genes, especially in regions containing quantitative elements, suggesting that these sites and the proteins that bind to them may have a role in enhancing gene expression. Such AT-rich binding sites, in quantitative regions, are present in genes coding for phaseolin (Bustos et al., 1989; Jordano et al., 1989; Guiltinan et al., 1989), pea legumin (Shirsat et al., 1989; Meakin and Gatehouse, 1991), helianthinin (Jordano et al., 1989), lectin (Jofuku et al., 1987; Lindstrom et al., 1990), phytohaemagglutinin (Riggs et al., 1989) and 19 kDa zein (Maier et al., 1990; Grasser et al., 1990).

The 55 bp AT-rich binding region of Pvu-β activates gene expression when ligated to a chimeric gene (Bustos et al., 1989). Similarly, the AT-rich binding region of LegA1 can activate expression of an otherwise inactive gene construct (Appendix II). These findings, coupled with their prevalence of AT-rich binding sites in quantitative elements, support the hypothesis that the AT-2 binding sites contain cis-acting elements that enhance Psa-v gene expression.

The presence of repeated cis-acting elements can have a greater than additive effect, due to cooperative binding or other interactions between the trans-acting factors. Synergistic enhancement of gene expression by the addition of extra elements has been observed with CaMV 35S enhancer (Kay et al., 1987) and ARE (Olive et al., 1990; Czarnecka et al., 1992). For example, anaerobically-induced expression of a chimeric gene construct containing 6 copies of the ARE was 12-fold higher than the gene with only one ARE (Olive et al., 1990). The second AT-2 element may have a similar role in the upstream quantitative element, such that the "major" and "minor" quantitative elements may be of equal effectiveness on their own. On the other hand, the presence of other specific binding sites in the major quantitative region (Figure 3.17), suggests that the greater enhancement by this region might be due to the presence of additional cis-acting elements.
SEQUENCE SPECIFICITY OF AT-RICH BINDING PROTEINS

Despite a lack of primary sequence similarity, AT-rich binding sequences from different seed-specific and other seed-expressed genes may be bound by similar or the same proteins; for example, AT-rich binding site sequences from the 5' region of genes encoding *P. vulgaris* phaseolin and phytohaemagglutinin (lectin); soybean lectin, and Kunitz trypsin inhibitor; and a sunflower helianthinin; can differentially cross-compete for binding proteins isolated from *P. vulgaris* (Riggs *et al.*, 1989) and sunflower seed (Jordano *et al.*, 1989). These results indicate firstly, that homologous proteins, that bind to AT-rich sequences with similar binding affinities and functional roles, may be present in most plant species. Secondly, the AT-rich sequences may primarily be recognised by an aspect of B-DNA configuration, such as specific configuration of the minor groove of the DNA helix (Solomon *et al.*, 1986) rather than recognising a specific nucleotide sequence.

Not all proteins appear to have loose binding affinities and so these may represent a different class of AT-rich binding proteins. The AT-2 binding region in *Psa-v*, was unable to compete with nuclear proteins binding to an AT-rich DNA-binding sequence in the *LegA1* legumin gene (Fig 3.17). Thus, AT-2 did not bind to the legumin fragment (or AT-2 had a much higher affinity for the legumin binding sites). This specificity of binding suggests that AT-2 may belong to this second class of AT-rich binding proteins. The binding protein, NAT-2, binds to 5' regions in nodule-specific genes (Jensen *et al.*, 1988; Metz *et al.*, 1988; Jacobsen *et al.*, 1990) may also belong to such a class of proteins. Although it recognises an AT-rich site, minor changes in the NAT-2 site, (from the soybean N23 nodulin gene) even some single base transversions from A→T resulted in a significant decrease in binding of NAT-2 (Jacobsen *et al.*, 1990).

THE AT-2 CONSENSUS SEQUENCE RESEMBLES BINDING SITES IN OTHER PLANT GENES

Of the different AT-rich binding sites in seed-storage protein genes, the AT-2 binding site in *Psa-v* most closely resembles the opposite strand of the SEF1 consensus (Allen *et al.*, 1989). Curiously, the AT-2 site also resembles the AT-rich binding site in box IV of the light-responsive promoter of pea *rbcS-3A* (Lam *et al.*, 1990). The AT-2 site is also similar to the consensus sequence of two AT-rich sites, bound by P-CAT, in an osmotic stress-responsive *Ppc1* gene from the common ice plant, *Mesembryanthemum crystallinium* (Cushman and Bohnert, 1992). The insert below shows the alignment of the AT-2 consensus sequence with these other AT-rich binding sequences, two alignments are
shown for box IV. (Bases corresponding to the 10 bp repeats in the AT-2 site are highlighted, bases that are not conserved within the respective binding sites are in lower case, lines represent identity with the AT-2 sequence).

<table>
<thead>
<tr>
<th>Comparison of the consensus sequence that binds AT-2 and other AT-rich binding sites</th>
</tr>
</thead>
<tbody>
<tr>
<td>aAATAAATAT</td>
</tr>
<tr>
<td>aACGGATAATAAATAGATAaATAAT</td>
</tr>
<tr>
<td>AATAGATAAAATAaAACATT</td>
</tr>
<tr>
<td>AATAGATAAAATAaAACATT</td>
</tr>
<tr>
<td>AAaTAACaAgTTTt</td>
</tr>
</tbody>
</table>

**AT-2 does not resemble High Mobility Group binding proteins**

Some nuclear proteins that bind AT-rich sites in plant genes resemble mammalian HMG (high mobility group) proteins. These include nuclear proteins from embryo-specific plant genes (Grasser et al., 1990; Maier et al., 1990; Pedersen et al., 1991; Goupil et al., 1992) as well as other plant genes (de Bruijn et al., 1989; Jacobsen et al., 1990; Czarnecka et al., 1992). HMG proteins are thought to play a role in activating transcription by relaxing the tertiary structure of DNA, by interfering with nucleosome positioning (Waga et al., 1989), or as components of the nuclear scaffold (Solomon et al., 1986), or by inducing bends in the DNA (Giese et al., 1992). HMG proteins are extractable in 0.35 M NaCl, and are also soluble in 2-5% trichloroacetic acid (Bustin et al., 1990). The HMG-like LAT-1 is also undenatured by heat treatment at 90°C for 10 min (Jacobsen et al., 1990). Whilst some AT-2 DNA-binding activity was retained after heat treatment, AT-2 was not soluble in 2% trichloroacetic acid, (data not shown) and had relatively low mobility, which indicates that it is probably not an HMG-like protein.

To summarise:

Many nuclear proteins were detected binding to the 5' region of Psa-v. The abundance and activity of these proteins varied during seed development and between tissues, suggesting a dynamic and complex interplay between these putative transcription factors and the elements to which they bind during the regulation of vicilin gene expression.
CHAPTER 4. FUNCTIONAL ANALYSIS OF CONSERVED DNA MOTIFS AND PROTEIN BINDING SITES IN THE PROMOTER OF PSA-V.

4.1 Introduction

Promoter analysis, (5' deletion), of Psa-v showed that seed-specific expression was regulated by DNA sequences downstream of position -306 (Chapter 2). In Chapter 3 it was shown that nuclear proteins, termed TEMP-1, TEMP-2 and TEMP-3, bound within the seed-specific region of Psa-v. The binding activity of these proteins correlated with the temporally regulated transcription and translation of the Psa-v gene during seed development. Furthermore, the leaf extract contained a factor that repressed the binding activity of TEMP-1. These results suggest that these binding proteins may act as transcriptional activators of this seed-specific and temporally regulated vicilin gene.

It has been demonstrated for some genes expressed in the grains of cereals (Guiltingan et al., 1990; Lohmer et al., 1991; Roth et al., 1991; Marcotte et al., 1992; Schmidt et al., 1992) and other plant genes (Green et al., 1988; Davis et al., 1990; Olive et al., 1991; Czarnecka et al., 1992; Sarokin and Chua, 1992) that DNA binding sites and the proteins that bind to them can transactivate gene expression. Nevertheless, it has also been demonstrated that some DNA sequences in the 5' regions of plant genes that bind nuclear proteins can be deleted or mutated without having a discernible effect upon gene expression, making it difficult to assign a functional role to such DNA sequences. This second class of binding sites has been detected in seed storage protein genes (SEF1, Allen et al., 1989; SEF3, SEF4, Fujiwara and Beachy, 1992) as well as in a vegetative storage gene (AT-rich site in patatin, Liu et al., 1990) and in genes involved in nodulation (NAT-2 site in nodulin genes, Stougaard et al., 1990; Jorgensen et al., 1991). The proteins binding to these DNA sequences are tissue-specific and temporally regulated, (Allen et al., 1989; Jacobsen et al., 1990; Lessard et al., 1991). It is, therefore, important to ascertain whether the TEMP-1, -2 and -3 binding sites act as regulatory elements in the seed-specific region of Psa-v.

Not all regulatory elements have been shown to contain sequences that are recognised by nuclear proteins. The CATGCATG motif in legume genes (Dickinson et al., 1988) does not appear to be bound by specific nuclear proteins (Jofuku et al., 1987; Allen et al., 1989; Bustos et al., 1989; Lessard et
106

al., 1991; Meakin and Gatehouse, 1991; Fujiwara and Beachy, 1992; Kawagoe and Murai, 1992). A multimerized CATGCATG oligo also failed to bind nuclear proteins (Riggs et al., 1989). Nevertheless, recent reports (Baumlein et al., 1992; Fujiwara and Beachy, 1992; Lelievre et al., 1992) have shown that CATGCATG motifs are bi-functional elements that contribute to high level expression in seeds and repress (to some extent) expression in non-seed tissues. Thus, not all regulatory elements can be located using DNA-protein binding assays.

This chapter describes experiments designed to investigate the functional significance of the TEMP-1, -2 and -3 binding sites, the role of the CATGCATG motifs and other conserved sequences in this gene and to locate (other) cis-acting elements within the seed-specific region of Psa-v. The seed-specific region, from -306 to +39, was used to make three sets of gene constructs, containing deletions, internal mutations and insertions (of additional binding sites conserved sequences) within the 5' region. These mutated 5' regions were ligated to a reporter gene encoding β-glucuronidase (GUS).

A transient expression system was developed in which the chimeric GUS genes were assayed, together with a modified luciferase gene (Bruce and Quail, 1990) which was used as an internal control. The genes were delivered by particle bombardment to pea leaves, petals, stems and developing cotyledons. The regulation of the transient expression of these constructs in these organs was similar to the seed-specific and temporally regulated expression of vicilin genes observed in vivo (Beach et al., 1985). This indicated that this transient system was suitable to investigate the seed-specific and temporal regulation of Psa-v. This method of gene analysis is relatively fast and has the additional advantage that gene expression was assayed in pea organs, rather than in heterologous tobacco host plants.

Analysis of the Psa-v-GUS gene constructs identified several regulatory regions required for high level gene expression. The results suggest that the binding sites, particularly the TEMP-1 sites, have a major role in transactivation of Psa-v during seed development. The results also indicate that a CATGCATG motif, tATGCATG and a similar motif, ggTGCATG, contribute to the positive regulation of Psa-v gene expression in seeds and may repress expression in other organs.
4.2 Materials and Methods

4.2.1 Plasmids

\textit{PSA-v-GUS CASSETTE}

An initial construct, \( \Delta-306/+39 \) \textit{Psa-v-GUS -Psa-v}, otherwise referred to as pQ20, was constructed by ligating the \( \beta \)-glucuronidase (GUS) coding region from pRAJ275 (Jefferson, 1987) as an \textbf{Nco I - Eco RI} fragment into the \textbf{Nco I} and \textbf{Xho I} sites of pCW71 (\( \Delta-306 \), Chapter 2). Plasmid pQ20, contained the 5' region of \textit{Psa-v} from -306 to +39, the GUS coding region and the 3' region of \textit{Psa-v} in a pUC118 vector. This gene construct was then used to generate mutants, to be used to identify regulatory elements within the seed-specific region of \textit{Psa-v}.

\textbf{5' DELETIONS}

A set of gene constructs were made, containing 5' deletions of the \textit{Psa-v} (Higgins \textit{et al.}, 1988) seed-specific region (-306/+39), ligated to the GUS coding region and the 3' region from \textit{Psa-v}. Plasmid pQ20 was digested with \textbf{Pst I} at the 5' polylinker site, followed by digestion with \textbf{Bal 31} for 5 to 7 minutes. DNA was end-filled using DNA polymerase I (Klenow fragment), and fragments ranging from 350 bp to 20 bp were released by digestion with \textbf{Nco I}. The fragments were separated on a polyacrylamide gel and then eluted from the gel slices (Chapter 4.2.3). The fragments were ligated to a \textbf{Bam HI} (end-filled) \textbf{Nco I} fragment of pQ20 (which contained only the coding region and 3' end of the gene plus pUC118). This resulted in a nest of constructs containing the \textit{Psa-v} 5' region deleted at positions ranging from +9 to -287, which were determined by sequencing.

\textbf{3' DELETIONS}

A set of constructs containing 3' nested deletions of the \textit{Psa-v} 5' region were also made. Plasmid pQ20 was linearised, at +39, by digestion with \textbf{Nco I}. Aliquots of the linear plasmid were digested with \textbf{Bal 31} for 2 to 7 minutes, end-filled using DNA polymerase I (Klenow fragment), digested with \textbf{Nar I}, and then the DNA fragments were separated using polyacrylamide gel electrophoresis, and eluted. The 3' deletion fragments were ligated into pUC118 between the \textbf{Nar I} and \textbf{Hinc II} sites. Clones containing between -306/+13 to -306/-230 were isolated and identified by sequencing.
INTERNAL MUTATIONS.

Internal mutations within the seed-specific region (symbolised by Σ, for substitution) were made by ligating fragments from the 3' deletion clones into appropriate 5' deletion clones. By using different restriction sites present in the polylinkers of the 3' and 5' clones, intervening sequences could be generated to replace sequences, ranging in length from 12 to 58 bp, between the 3'/5' junction with no more than 2 bp added or deleted. Constructs pQ-ΣC to pQ-ΣI were made in this fashion, resulting in (pQ-) C Σ-230/-194, D Σ-198/-175, E Σ-186/-150, F Σ-153/-117, G Σ-129/-96, H Σ-98/-79, and I Σ-88/-73. Constructs A Σ-306/-231 and B Σ-307/-272 are simply 5' deletions, since 3' clones smaller than A3'c-230 (Psa-v from -306 to -230) were not isolated. Construct K Σ-61/-4 was also constructed differently, because of the lack of an appropriate A3' clone. The Nsp I fragment from clone pQ-Δ3'k+13 containing Psa-v sequences between -306 and -61 was ligated into the Sph I site of pQ-A5', producing a construct mutated between -61/-4 of Psa-v, but with a 32 bp (3 helical turns) less intervening sequence. A construct with mutations spanning -73/-61 was not isolated. The locations of the base-pair substitutions, and of the conserved sequence within the Psa-v 5' region of these constructs, are indicated in figure 4.1. The precise nucleotide changes are shown in figure 3.2.

DUPLICATIONS AND REVERSIONS

Gene constructs containing duplicated or inverted regions of the 5' region of Psa-v were made by ligation of an Xba I fragment of plasmid pQ-Δ3'i-88 (see 3' deletions, above) containing sequences between -306 to -88 into the Xba I site of plasmids containing different 5' deletions. Ligation of the -306/-88 fragment into plasmid pQA5'i-73, in forward and reverse orientation, resulted in plasmids pQ-Σ1-88/-73 (above) and pQ-R-88/-73, respectively.

Similarly, ligation of the Xba I fragment in forward and reverse orientation into pQ-Δ5'e-150 resulted in plasmids pQ-I-88/-150 and pQ-R-88/-150, respectively. The Xba I fragment was also inserted into the Xba I site of pQ-Δ5'c-194, generating plasmid pQ-I-88/-194. (The locations of the duplicated and inverted sequences are shown in Figure 4.8).

To make it easier to identify between the different deletion, mutation, duplication and reversion gene constructs, these are referred to in the remainder of the text without their pQ- prefix.
Figure 4.1
Mutations in the 5’ sequence of \textit{Psa-v}.

The wildtype gene construct (wt) contains -306 to +39 of \textit{Psa-v}, ligated to the GUS coding region and the \textit{Psa-v} 3' region. A series of constructs containing mutations in the 5' region were as follows: \(\Sigma A\) -306/-272, \(\Sigma B\) -306/-231, \(\Sigma C\) -230/-194, \(\Sigma D\) -198/-175, \(\Sigma E\) -186/-150, \(\Sigma F\) -153/-117, \(\Sigma G\) -129/-96, \(\Sigma H\) -98/-79, \(\Sigma I\) -88/-73, \(\Sigma K\) -61/-4, indicated as A to K, respectively. The locations of the substituted nucleotides are indicated by crosshatching. The conserved and repeated DNA sequences within the 5' region of \textit{Psa-v} are shown; from right to left: TATA box (T), CAAT box (C), CATGCATG elements (\(\uparrow\), sequence tATGCATG, \(\downarrow\), sequence ggTGCATG), vicilin box I (overlined, containing a CACA box, \(\nearrow\) and an octanucleotide box, \(\searrow\)), vicilin box II (\(\swarrow\)) and a palindrome (\(\leftrightarrow\)). Positions are relative to the start site of transcription. The sequences of the 5’ regions of these constructs are shown in Figure 3.2.
The construct a-2500/wt-272 region of Paa-1 ligated to the CaMV promoter was kindly provided by Marlene and Antanasius Marka.

This construct contained the upstream promoter limited to the enhancer's internal region and was used as a positive control. Another construct, containing BAC genomic DNA, was used as a negative control. This construct contained the upstream promoter limited to the enhancer's internal region. This construct was then constructed to function in transgenic tobacco plants, particularly experimental plants containing a NOS 3' region.

4.3.2 Particle Bombardment

Pea cotyledons were grown in the greenhouse and were harvested from mature pea plants. The seeds from the pea plants were removed, and the cotyledons were separated. Pea cotyledons (inner side facing up), and slices of potato, potato leaves and stems were arranged in a circle (25 mm in diameter) on MSB agar plates (see Chapter 2) containing 150 μg/ml tunicamycin to prevent contamination by bacteria. Filter paper circles were
POSITIVE AND NEGATIVE CONTROL CONSTRUCTS

The construct \( \Delta -2500 \text{vic-GUS-35S} \), contains the full 5' region of \( Psa-v \) ligated to the GUS coding region and the 3' region of the CaMV 35S gene. This construct (formerly termed EcoRIvic2.5kbHind III12-merGUS.ter35sHind III) was kindly provided by Marjorie and Antonius Matzke.

The construct p35SGV was kindly provided by Paul Howell. This construct (formerly termed PA1, and renamed to avoid confusion with pea albumin I) was used as a positive control for GUS expression in bombarded organs. The 5' region of \( Psa-v \) in pQ20 was removed and replaced with a fragment containing the promoter of the CaMV 35S gene (isolated from pBl121, Clontech).

Another construct, p35SGN100, was also used as a positive control. This construct, kindly provided by Danny Llewellyn, contains the 35S promoter region (-420 to +1), ligated to the GUS coding region and nopaline synthase (NOS) 3' region.

Plasmid pAHCl8 (Bruce and Quail, 1990) was used as an internal standard. This construct contains the ubiquitin promoter ligated to the (enhancer) intron from the maize \( Adh1 \), followed by the gene coding region of the firefly luciferase gene and the NOS 3' region. (Although this construct was initially constructed to function in monocotyledonous plants, preliminary experiments established that it conferred moderate levels of expression within pea tissues).

4.2.2 Particle bombardment

PLANT MATERIAL AND CULTURE

Peas (\( Pisum sativum \) L., cv. Greenfeast) were grown in the glasshouse as described in Chapter 3. Pea flowers, seeds, pods, leaves and stems were harvested from maturing pea plants. The testa from the pea seeds were removed, and the cotyledons were separated. Pea cotyledons (inner sides facing up), and slices of petals, pods, leaves and stems were arranged in a circle (25 mm in diameter) on MSO agar plates (see Chapter 2) containing 150 \( \mu \)g/ml timentin to prevent contamination by bacteria. Filter paper circles were...
also moistened with MSO liquid medium, and these were placed in the lid of the petri dishes to maintain high humidity. After transfection by bombardment (see below) the cotyledons were turned over, so that the shot surface did not dry out, and these and the other tissues were incubated at 25°C for 48 hours. The tissues were either assayed for GUS activity by fluorometric assay or by histochemical staining. The tissues that were assayed fluorometrically were frozen in liquid nitrogen and stored at -80°C until extraction. To reduce dilution of activity by the presence of an excess of non-transfected tissue, the face of the cotyledons that had been shot with the tungsten particles was removed from the remainder of the cotyledon with a razor, immediately prior to freezing.

DNA PRECIPITATION

Plasmid DNA was precipitated onto tungsten particles essentially as described by Bruce et al. (1989). Each test construct (5 µg) was mixed with plasmid pAHC18 (1 µg) in a volume of 12 µl. Tungsten particles (25 µl at 100 µg/µl in 50% (v/v) glycerol) were added to the DNA mix, followed by CaCl2 (25 µl at 2500 mM) and spermidine (free base, 10 µl at 1000 mM). The mixture was shaken gently for 4 min, centrifuged (10000 x g) for ten seconds, then 32 µl of supernatant was discarded. The pellet was resuspended briefly by sonication, immediately prior to bombardment of the tissue.

GUNPOWDER-DRIVEN GUN

Initial experiments were done using a MAB Gene Gun, essentially as described in Bruce et al. (1989); Bruce and Quail (1990). The MAB Gun employed a gunpowder charge to accelerate a .22" calibre nylon macroprojectile loaded with four µl of the DNA-coated tungsten particles (containing 0.5 µg test plasmid DNA) towards a steel stopping plate with a 1.5 µm hole in the centre. The tungsten particles passed though the hole at high speed and impacted with the target tissue, which was protected by a wire mesh, 8 cm below, in an evacuated chamber (85 kPa).

HELIUM GUN

A Particle Inflow Gun (PIG, Finer et al., 1992), which employed helium, rather than a gunpowder charge, became available after the first experiments had been done using the MAB gene gun. The remaining experiments were done with the PIG gun because it was cleaner, faster, and cheaper to operate (and considerably quieter), damaged the tissues less, and (perhaps because there
was less wounding) resulted in 5-20 fold greater levels of GUS activity in transfected tissue than the gunpowder-driven MAB gun. With this system, 6 µl of DNA-coated tungsten particles were loaded into the centre of a filter of a Swinney unit (Millipore, USA) and the unit was twisted into position in the gun chamber. The particles were accelerated by a short burst of helium gas (at 690 kPa) into the evacuated chamber (85 kPa) onto the target tissue, which was protected by a wire mesh, 10 cm below.

4.2.3 Luciferase assays

Luciferase assays were essentially as described by Lanahan et al. (1992). The frozen samples were homogenised, using Eppendorf grinders (fused and etched 1 ml glass pipettes) and a little acid-washed sand in 600 µl of luciferase extraction buffer (200 mM KPO4, 1% (v/v) Triton X-100, 5 mM DTT, pH7.5). The homogenates were centrifuged (10 000 x g) for 10 min and the supernatants were retained for both luciferase and GUS enzyme assays.

For the luciferase assays, 100 µl samples of the extract were placed into disposable glass test tubes and then 100 µl of 2 x luciferase assay buffer (2 x LAB: 60 mM Tris-SO4, 20 mM MgCl2, 20 mM DTT, 1 mM EDTA, pH 7.7), with 2 mM ATP was added immediately before placing it in the luminometer (Lumat LB9501, Berthold, Germany). The machine was programmed to inject 100 µl of luciferin solution (1 x LAB containing 1 mM luciferin), and then to count the photons emitted for 10 seconds. Typical values for luciferase activity were between 1 000 and 30 000 luciferase units (LU). Background activity, from extracts from tissue that had been bombarded by 'blank' tungsten particles, was between 100 to 200 LU.

4.2.4 GUS assays

HISTOCHEMICAL STAINING

The tissues were immersed for 48 h at 37°C, in the dark, in a solution containing 1 mM 5-bromo-4-chloro-3-indolyl-β-glucuronic acid (X-Gluc), 50 mM sodium phosphate buffer pH 7.0, 1 mM EDTA, 0.5 mM potassium ferricyanide, 0.5 mM potassium ferrocyanide and 20% (v/v) methanol. Subsequently, chlorophyll was removed from green tissues by stepwise incubations in 50% (v/v) and 65% (v/v) ethanol for 30 min and then in 70% (v/v) ethanol for 48 h.
FLUOROMETRIC GUS ASSAY

GUS activity was quantified using the fluorescent 4-methlyumbelliferyl β-D-glucuronide (MUG) substrate (Jefferson et al., 1987). Protein extract (50 µl) was incubated with 50 µl of GUS assay buffer (2 mM MUG in 50 mM sodium phosphate buffer, 10 mM EDTA, 20% (v/v) methanol, 0.1% (v/v) Triton X-100, 0.1% (v/v) lauryl sarkosine, 10 mM β-mercaptoethanol, pH 7.0) at 37°C, in the dark, for 20 hours. The reaction was terminated by adding Na₂CO₃ to a final concentration of 20 mM. Fluorescence was measured using an excitation wavelength of 360 nm, and an emission wavelength of 455 nm. Care was taken to ensure that all assays were done within a linear range of detection of 4-methyl umbelliferone.

DATA ANALYSIS

Background luminescence activity, produced by tissues bombarded with "blank" tungsten particles, was subtracted from the luciferase activity values for the other plasmids. The rate of GUS activity in extracts from the transfected tissue was normalised by the luciferase expression in the same extracts. Normalised GUS activity was calculated as the GUS activity (pmol 4MU/hour for 50 µl extract) divided by the relative luciferase activity (luciferase units/10 sec/100 µl extract). Although gene expression in the cotyledons varied between experiments, probably due to environmental variations between plants, the relative expression between different constructs remained constant. Therefore, the seed data are presented as a percentage of that obtained with the construct Δ-306-GUS.

Statistical comparisons of results from different constructs were done using Student's t test. The probability of differences in activity between constructs was considered significant at P<0.10. In most cases, however, and unless otherwise stated in the text, significant differences between data were at the 5% level (P<0.05).
4.3 Results

4.3.1 Assessment of specificity and temporal regulation of activity in transfected organs

To determine whether transient expression of gene constructs containing the 5' region of *Psa-v* maintained tissue-specific and temporal regulation, different pea organs were transfected using particle bombardment.

The construct A-2500vic-GUS-35S contains the full 5' region of *Psa-v*. It has been shown that this construct was expressed highly in the embryonic tissue of seed from stably transformed tobacco. The GUS activity of this construct was also temporally regulated, peaking at 18-19 DAF (approximately 43 300 pmol/mg protein/min, A. Matzke and M. Matzke, pers. comm.). Since this gene construct was both correctly regulated, and expressed at high levels in transgenic plants, it was used in the initial bombardment experiments to test whether organ-specific and temporal regulation by the 5' region of *Psa-v* was maintained. Histochemically stained pea cotyledons that had been bombarded with A-2500vic-GUS-35S had intense blue spots, indicating a high level of gene expression (Figure 4.2). Smaller, much less intense spots were occasionally observed on pea petals, pods, leaves and stems transfected with this construct. Hence the transient expression conferred by the *Psa-v* 5' region was essentially, although not completely, seed-specific.

The majority of the *Psa-v* constructs that were assayed also contained the 3' region of *Psa-v*. Since the 3' flanking regions of some plant and animal genes contain elements that may affect tissue-specific and temporal regulation (Choi and Engel, 1986; Fischer and Maniatis, 1986; Dietrich et al., 1992), plasmid p35S-G-V, containing the 35S promoter and the *Psa-v* 3' region, was also tested for organ-specific expression. This construct was expressed to a high level in pea leaves and petals, and was expressed at lower levels in pea seeds, pods and stems (Figure 4.2). Thus the low activity of various *Psa-v* constructs in the vegetative organs, was not due to the presence of the 3' region of vicilin.

Transient expression of the chimeric A-2500vic-GUS-35S gene also correlated with the temporal regulation of vicilin gene expression in peas. Expression of the gene occurred in cotyledons between 10 and 16 DAF, peaking at day 13 (Figure 4.3), which corresponded to the stages when vicilin mRNA levels are increasing during pea seed development (Figure 3.4). This contrasted with the
Figure 4.2
Organ-specific expression of a *Psa-v*-GUS construct.

Transient GUS activity in pea cotyledons, pods, petals, stems and leaves that were transfected, using particle bombardment. Organs in the left panels were bombarded with construct Δ-2500vic-GUS-35S, which contains the entire 5' region of *Psa-v*, ligated to the GUS coding region and the 3' region of the CaMV 35S gene. Those in the right panels were bombarded with p35SGV, which contains the CaMV promoter, ligated to the GUS coding region and the 3' region of the *Psa-v* gene.
Figure 4.3
Temporal expression of a *Psa*-v-GUS gene construct.

Transient GUS activity in pea seeds that were transfected, using particle bombardment, with \( \Delta \)-2500vic-GUS-35S (top and middle panels) or p35SGN100 (bottom panel). The developmental stages of the cotyledons at the time of bombardment (days after flowering) are indicated (from the right).
Days After Flowering

Days After Flowering
expression of p35SGN100 which was both less intense than that conferred by the *Psa-v* construct (Figure 4.3), and which peaked in older seeds, after 17 DAF (not shown).

### 4.3.2 Four regulatory DNA elements in the seed-specific region of *Psa-v* were identified by analysis of 5′ deletion constructs.

The expression of a set of *Psa-v*-GUS constructs, with 5′ deletions downstream of position -306, was analysed using particle bombardment of pea cotyledons (11-16 DAF). GUS activity in pea cotyledons transfected with four of these gene constructs, Δ-306-GUS, Δ-117-GUS, Δ-79-GUS and Δ-4-GUS bp of the 5′ region is shown in Figure 4.4. (The organs in Figure 4.2 and 4.3, with the exception of the petals, were bombarded using the MAB (gunpowder-driven) gun. Bombardment by the PIG (helium) gun however, which arrived after the experiments described in 4.2.1 had been done, resulted in much greater GUS activity, which can be observed in Figure 4.4).

Quantification of GUS activity in cotyledons bombarded with 5′ deletion constructs is shown in Figure 4.5. No significant change from Δ-306-GUS expression was detected (P>0.10), until the deletion to -117 (P<0.05), suggesting that there is a positive element within -150/-118. Deletion of the adjacent region -117/-89 reduced expression further to approximately half of Δ-117-GUS activity. No significant change from this level of activity was noted until the deletion of 12 bp between -73 and -62, which reduced gene expression to background levels.

The GUS activity in leaves, stems and petals bombarded with selected 5′ deletion constructs was also quantified (Figure 4.6). The level of GUS activity was much lower in these organs than in seed (0.1-1.3%), but was significantly above background levels. Two exceptions were the Δ-117-GUS and Δ-4-GUS constructs.

Although deletion of sequences between -306 to -195 did not have a significant effect on GUS activity in cotyledons, deletion of this region resulted in a significant increase in GUS activity in petals and leaves, suggesting that there was negative element between -306 and -195. Deletion of the adjacent region between -194 and -118 significantly reduced activity, almost to background levels, in these non-seed organs. Subsequent deletion of -117 to
Figure 4.4
Quantitative changes of GUS activity in pea cotyledons bombarded with 5' deletion Psa-ν-GUS gene constructs.

Pea cotyledons, harvested between 11-16 DAF, were transfected with (A) Δ-306-GUS, (B) Δ-117-GUS, (C) Δ-79-GUS and (D) Δ-4-GUS. See Figure 4.5A for details of the 5' regions of the gene constructs.
Figure 4.5

5' deletion analysis of the seed-specific region of the *Psa-v* gene.

A. A set of *Psa-v*-GUS gene constructs, each containing 5' deletions of the *Psa-v* 5' region, ligated to the GUS coding region and the 3' flanking region of *Psa-v*. The location of the 5' endpoints and the conserved sequences are shown. The symbols are described in the legend of Figure 4.1.

B. Each column and line in the histogram represents the mean activity of 5-14 pea cotyledons, from 1-3 separate experiments, plus standard error. Cotyledons, approximately 11 to 16 DAF, were bombarded with *Psa-v*-GUS plasmids co-precipitated with a luciferase (control) plasmid onto tungsten particles. GUS activities are shown as a percentage of the normalised GUS expression of the Δ-306-GUS construct.
A.

B. 140

% Activity (± s.e.)

5' Deletions
Figure 4.6
GUS activities in petals, leaves and stems bombarded with 5' deletion
Psa-v-GUS gene constructs.

The relative GUS activity in extracts from pea petals (top panel), leaves (middle
panel) and stems (bottom panel), after bombardment by 5' deletion Psa-v-GUS
plasmids, or negative control (luciferase only) or positive control (p35SGN100)
plasmids. Each column and line in the histogram represents the mean (plus
standard error) of 2-7 samples, from two separate experiments. Normalised
GUS activity, as described in methods is expressed as $10^3 x \text{pmol} \ 4\text{MU/h/relative luciferase activity}$. Average expression of Δ-306-GUS in
extracts from bombarded pea cotyledons from these experiments was $150 \times 10^3 \pm 25 \text{pmol 4MU/hour/ relative luciferase activity}$. 
-90 increased GUS activity in petals and leaves, indicating removal of a second negative element. In contrast, deletion of this region reduced the level of GUS activity in seeds, which suggests that this region has dual functions; to enhance seed-specific activity and concomitantly inhibit it in other organs. Further deletions had no significant effect on activity until the deletion between -73 and -4, which reduced the expression in petals, leaves and stems to background levels. Although the variation in GUS expression in stems was too large to measure significant changes in activity for constructs between -306 to -73, the average pattern of expression of the 5' deletion mutants was similar to that observed in leaf and petals, suggesting that the regulatory elements in Psa-v had a similar effect on gene expression in stems as well as in leaves and petals.

Organ-specific expression of the Psa-v gene constructs was screened by histochemical staining of transfected leaves and stems for GUS activity. The histochemical method was found to be as sensitive as the fluorometric GUS assays, and was sufficient to show that there was no marked change in the organ-specific expression (data not shown).

4.3.3 Transient expression of gene constructs with mutations within the region -306 to +39 of Psa-v.

Promoter analysis using simple 5' deletion of gene promoters constructs may not detect all the regulatory elements. Therefore, gene constructs containing internal mutations within -306 to +39 of Psa-v (Figures 4.1, 4.7) were used to locate further regulatory elements. Three regions, including two identified by 5' deletion analysis (Figure 4.5) were shown to have a major role in regulating Psa-v gene expression. Mutation of sequences between -153/-117, -129/-96, and -61/-4 all significantly (P<0.05) reduced GUS activity to about 25% of that conferred by the intact -306-GUS construct (Figure 4.7). The 13 bp between -73/-61, were not mutated within any of the internal mutation constructs and so the relative contribution of this region to Psa-v expression was not determined.

The mutation of three other regions, also significantly (0.05<P<0.10) reduced gene expression; mutation within -186/-150, -98/-79 and -88/-73 all decreased expression to about half that conferred by the wild-type (Δ-306-GUS) gene (Figure 4.7). The expression of constructs ΣA-306/-272, ΣB-306/-231 and ΣD-198/-175 did not differ significantly from Δ-306-GUS (P>0.10). Mutation of
Figure 4.7
Mutational analysis of the *Psa-v* promoter.

A. A set of gene constructs consisting of internal mutations in the *Psa-v* 5' region which were ligated to the GUS coding region and the 3' flanking region of *Psa-v*. The location of the mutations and the conserved sequences are shown. The symbols are described in the legend of Figure 4.1.

B. Each column and line in the histogram represents the mean activity in extracts of 5-14 individual pea cotyledons, from up to 3 separate experiments, plus standard error. Cotyledons, approximately 11 to 16 DAF, were bombarded with *Psa-v*-GUS plasmids co-precipitated with a luciferase (control) plasmid onto tungsten particles. GUS activities are shown as a percentage of the normalised GUS expression of the Δ-306-GUS construct.
<table>
<thead>
<tr>
<th></th>
<th>wt</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
<th>K</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**B.**

![Bar chart showing % Activity (+ s.e.)](chart_url)

**Legend:**
- **WT:** Wild Type
- **A - J:** Various internal mutations
- **Luciferase**
the sequence between -230/-194, however, increased expression by approximately 50%, indicating that this sequence contains a negative element.

4.3.4. Functional analysis of the TEMP-1 binding regions

Competition experiments showed that three regions of Psa-v, between -186/-150, -153/-117 and -129/-96 are important for binding of TEMP-1 (Chapter 3). Each of these regions contains the consensus sequence TTttTtcAtTTTgacacat, located at -186/-165 (in reverse orientation), -144/-125, and -121/-102 (Figure 3.13) and so it is proposed that TEMP-1 binds to each of these sequences. (Although it has not been shown formally that each of these overlapping regions contains an individual binding site for TEMP-1, for simplicity they are referred to and depicted as separate TEMP-1 binding regions.)

The previous experiments (Figures 4.5 and 4.7) showed by deletion and mutation that each of these three regions was required for high level expression of the Psa-v gene, especially the proximal sequences between -153/-117 and -129/-96 (Figures 4.5, 4.7). These two regions also contain the conserved sequence vicilin box I, located between -131 and -99. The weaker quantitative element, -186/-150, contains another vicilin-specific motif, vicilin box II. Since the proximal TEMP-1 consensus sequences appeared to have a greater effect than the inverted distal repeat, the orientation and/or position of the sites within the 5' region may influence gene regulation. Therefore, a series of gene constructs were made; primarily to test whether the presence of extra copies of the regulatory regions containing these TEMP-1/vicilin box sequences affected gene expression, and also to examine the effect of orientation and position of the regions.

A DNA fragment, -306/-88, was inserted at different positions and orientations in various Psa-v-GUS constructs. This fragment contains all the TEMP-1 sites/vicilin boxes plus the sequence tATGCATG (between -96/-89). Addition of this fragment to \( \Delta \)-73-GUS, a construct that was expressed at low levels, did not significantly increase GUS activity (Figure 4.8; compare \( \Delta \)-88/-73 with \( \Delta \)-73-GUS). This suggested that these sites may not enhance activity in the absence of the CAAT box (-87/-72). Addition of -306/-88 in the opposite orientation, however, significantly (\( P<0.05 \)) enhanced gene expression and increased it to levels equivalent to that of the intact seed-specific region. The 15 bp between -88 and -73 in construct \( \Delta \)-306-GUS were replaced by only 14
Figure 4.8
Functional analysis of region -150 to -88 of *Psa-v*, containing sequences required for TEMP-1 binding.

A. The 5' region of the constructs, containing up to 6 TEMP-1 putative consensus binding sites, in forward or inverted orientations (semicircles above and below) are shown in the top panel. The approximate locations of vicilin box II (\(\nabla\)) and the octanucleotide box (\(\nabla\)), and CACA box (\(\nabla\)) of vicilin box I are indicated. Regions containing polylinker sequence are cross-hatched.

From top to bottom: fragments from Δ-300-GUS, Δ-73-GUS, Δ-89-GUS, ΣI-88/-73, RI-88/-73, Δ-150-GUS, I-88/-150, R-88/-150, Δ-194-GUS and I-88/-150 are depicted (as -306, -73, -89, -88/-73, R-88/-73, -150, I-88/-150, R-88/-150, -194 and I-88/-194 respectively)

B. Mean activity of 5-14 pea cotyledons transfected with the constructs shown in panel A. Data are from up to 3 separate experiments, plus standard error. Cotyledons, approximately 11 to 16 DAF, were bombarded with *Psa-v*-GUS gene constructs which were co-precipitated with a luciferase (control) plasmid onto tungsten particles. GUS activities are shown as a percentage of the normalised GUS expression of the Δ-306-GUS construct
A. 

The sequence between -306 and +1 was expressed to determine the levels of promoter activity. Deletion of the upstream elements resulted in increased activity by having no promoter elements thus increasing the levels of transcription. 

B. 

To examine the effect of the deletion, a construct was comprised of two pairs of levels to compare. The construct also contained a TEMPER1 binding site, which was tested for activity. Since the GUS activity of the WT was that of I-88/-194, it was used as a marker for the activity of the construct. 

The graph shows the percentage of activity (+ s.e.) for each construct. The constructs are labeled as follows: 

-306, -73, -89, -88/-73, R-88/-73, -150, I-88/-150, R-88/-150, -194, I-88/-194.
bp of "foreign" DNA in construct ΣI-88/-73. Thus it is possible that the ineffectiveness of the ΣI-88/-73 construct was due to the disrupted interaction of upstream elements within -306/-88 with those proximal to -73, by spacing and/or helical constraints. In addition, the reverse orientation of -306/-88 may have increased activity by having altered the relative spacing between elements; thus increasing the length of DNA available for looping out, permitting transcriptional enhancement.

The sequence between -306/-88 was also added to a construct that was highly expressed, to determine whether it could enhance already high levels of expression. Ligation of this fragment to Δ-150-GUS essentially produced a duplication of -150/-88, containing the proximal binding sites, at position -150. This created a construct with one reversed and four other TEMP-1 consensus sequences. This construct, I-88/-150, had 10-fold higher activity than Δ-306-GUS (figure 4.8). Thus, the two extra sites had a synergistic rather than an additive effect on the enhancement of gene expression. Addition of -306/-88 in the reverse orientation to Δ-150-GUS (generating R-88/-150) conferred only 2.5-fold enhancement of gene expression (compared to Δ-150-GUS). This construct also contains 5 TEMP-1 consensus sequences, but two are in the reverse orientation and these, plus another TEMP-1 site, are located further away from the proximal sites than in construct I-88/-150 (Figure 4.8).

To examine the effect of distance/position of the duplicated sites, the fragment -306/-88 was ligated to Δ-194-GUS, a construct which was expressed a similar levels to both Δ-306-GUS and Δ-150-GUS. The construct that was generated, I-88/-150, essentially contained a duplication of -194/-88, containing the three TEMP-1 sites, inserted at position -194. This duplication also significantly (P<0.05) enhanced GUS activity; about three-fold compared to both Δ-194-GUS and Δ-306-GUS (Figure 4.8). Since the GUS activity of I-88/-194 was less than that of I-88/-150, this again suggests that spacing or the orientation of the TEMP-1 binding sites relative to each other, and perhaps other, downstream elements, may moderate the level of transcriptional enhancement. Alternatively, the presence of an additional copy of the region between -194 and -150 may have quenched some of the enhancement, but this seems unlikely since this region acts as a weak positive element (see Figure 4.7).
As shown in Figure 4.3, the transient expression of Psa-v-GUS constructs in pea cotyledons appeared to be regulated temporally. Thus, not surprisingly, some of the observed variation in GUS activity for individual constructs correlated consistently with slight differences between the developmental maturity of the pea cotyledons (data not shown). Initial experiments, however, indicated that the TEMP-1 insertion constructs, I-88/-150 (Figure 4.9), I-88/-194 and R-88/-150, were expressed most highly in very young cotyledons, about 9-10 DAF, at a stage when all Psa-v-GUS constructs (except perhaps Δ-230/-194) had little or no activity.

4.4 Discussion

4.4.1 Transient expression of a Psa-v-GUS construct was temporally regulated and essentially seed-specific

The GUS activity in pea cotyledons transfected with the construct Δ-2500vic-GUS-35S was highest in developmental stages between 10-16 DAF (Figure 4.3), indicating that temporal regulatory control was maintained in this transient assay system. GUS activity was not restricted to the seed, however; low-level activity was detected in petals, pods, leaves and stems transfected with this and other Psa-v-GUS constructs. GUS activity was far greater in the seed (about 80-fold), so that although organ specificity of Psa-v expression was not absolute, it was essentially restricted to the seed. Thus, particle bombardment of pea tissues could be used to examine both temporal and seed-specific aspects of gene expression of the Psa-v gene.

The low-level activity in the vegetative and floral organs may reflect genuine low-level activity of Psa-v in vivo. One of the three classes of vicilin genes in pea, of which Psa-v may be a member, is transcribed at very low levels in pea leaves (Beach et al., 1985); and the cDNA clone, p15-84, from this gene class, was used as a probe to isolate the Psa-v gene (Higgins et al., 1988). Furthermore, western blot analysis detected very faint bands reacting to vicilin antibody when high levels of total protein from pea stems and leaves (but not roots), and from tobacco leaves transformed with the complete Psa-v gene (Δ-2500, Higgins et al., 1988), were assayed. Since the Δ-2500vic-GUS-35S 3' construct also conferred moderate levels of expression within pea petal and pod tissue (Figure 4.2), Psa-v may also be expressed at low levels at earlier
Extra TEMP-1 binding sites enhance expression in young cotyledons.

Pea cotyledons between 9 and 18 DAF, transfected with Δ-306-GUS, which contains -306 to +39 of *Psa-v*, or I-88/-150, which contains -306 to +39 of *Psa-v*, with an insertion of sequences -150 to -88 (containing two TEMP-1 consensus sequences, within vicilin box I) at position -88. (See Figure 4.8.)
Figure 4.10
Summary of Promoter Analysis of Psa-v

A. Conserved motifs within -306 to +1 of Psa-v. The symbols are described in the legend of Figure 4.1.

B. Location of DNA-protein binding sites between -306 and +1 of the Psa-v gene. TEMP-I (putative) consensus sequences are shown as dark semicircles in forward(•) and reverse (◦) orientation. TEMP-2 and TEMP-3 sites, both located in -61/-4 are represented by stippled and empty arches (□, ◊).

Sections C, D, and E show regulatory regions within the Psa-v promoter, that affected expression in pea seeds. These regions were defined by analysis of 3 types of gene constructs, as indicated. The number of plus and minus symbols indicates the relative contribution of each region. Region -73 to -61 was not assessed (nd) by the internal mutation series.

F. Location of regulatory regions controlling expression in non-seed organs. GUS expression was not quantified for all the 5' deletions, so that regions are less well-defined than for seed-related expression.
MOTIFS

A. Psav

![Motif Diagram]

B. BINDING

![Binding Diagram]

SEED-Expression

C. Deletions

![Deletions Diagram]

D. Insertions

![Insertions Diagram]

E. Mutations

![Mutations Diagram]

NON-SEED-Expression

F. Deletions

![Deletions Diagram]
stages of flower, fruit and seed development.

Some vicilin genes are transcribed at very low levels in leaves and perhaps other vegetative organs. One possible reason why vicilin protein is not readily detected in these organs is that vicilin, which is normally targeted to the storage vacuole, and is far less stable in lytic leaf vacuoles (Wandelt et al., 1992). Hence it is possible that the weak expression of the construct Δ-2500vic-GUS-35S in bombarded pea leaves and stems reflects genuine low level expression, perhaps due to greater stability of GUS protein, that was not readily detected using the CAT (Appendix II) and vicilin (Higgins et al., 1988 and Chapter 2) reporter genes. Alternatively, if, as was suggested by the presence of a specific DNA-binding-inhibitor in leaf tissue (Chapter 3), Psa-v expression is actively repressed in some tissues, the presence of multiple copies of the construct within the bombarded cells may titrate, and so dilute, complete and effective repression by the inhibitory factor.

4.4.2 The region -306 to +39 of Psa-v contains many regulatory elements

Internal mutations, insertion/duplications and 5' deletions of the first 306 bp of the Psa-v promoter were used to determine the location of regulatory elements in this seed-specific region. A summary of the elements localised using these constructs, the position of these relative to conserved DNA motifs, and regions recognised by nuclear proteins is presented in Figure 4.10.

Three positive elements were identified by 5' deletion analysis; located between -150/-118, -117/-89 and -73/-61 (Figure 4.10C). Each element contained a different conserved DNA motifs: an octanucleotide box; a CACA box and a CATGCATG elements; and a CATGCATG-like motif, respectively (discussed below).

Analysis of constructs containing internal mutations within the seed-specific region supported the results from the 5' deletion analysis and identified other elements within this region (Figure 4.10E). Three regions containing strong positive elements (++) were identified, two overlapping regions that correlated with those identified with the 5' deletion results; between -153/-117, -129/-96 and another positive element between -61/-4, which contained the TATA box.
A weak negative element (−), which had not been identified by 5' deletion analysis, was located between -230/-194. In addition, weak positive elements (+) were located in the regions -186/-150, -98/-79 and -88/-73. Since the sequences between -98/-79 and -88/-73 overlap, they may contain a common element between -88/-79. But (5') deletion from -89 to -73 had no significant effect on gene expression, whilst deletion from -117 to -89, and from -73 to -61 both significantly reduced expression. Thus, it is more likely that the sequences contain two separate elements, located between -98/-88 and -79/-73. It is noteworthy that these regions contain a CATGCATG-element and a CAAT box, respectively, (these and conserved motifs present in other regulatory elements are discussed below).

Expression of the gene constructs containing duplications of the regions required for TEMP-1 binding supported the results from the analysis of both the 5' deletion and internal mutation constructs that showed that the elements between -150/-88 had a major role in the enhancement of gene expression (Figure 4.1).

4.4.3 TEMP-1 binding regions confer high level expression in seeds

The principal positive elements identified between -150/-118 and -117/-89 each contain a region required for the binding of the nuclear protein TEMP-1. This strongly suggests that these TEMP-1 sites contain cis-acting elements that confer high-level expression within seeds. Gene constructs were made and tested to determine whether the presence of two or three extra TEMP-1 binding regions would affect expression. The duplication of the region -150/-88, containing both these binding sites, at position -150 in the seed-specific region conferred 10-fold higher GUS activity within bombarded pea cotyledons (Figure 4.8). This construct, I-88/-150, had four tandem TEMP-1 sites, near the transcription start site, and an inverted TEMP-1 site further upstream. Other constructs with five (R-88/-150) or six (I-88/-194) TEMP-1 sites, also enhanced gene expression above wildtype (Δ-306-GUS) levels (Figure 4.8). Thus, as well as correlating with high expression levels in seeds, the extra copies of the TEMP-1 binding regions appear to enhance expression in seeds. This supports the proposal that transactivation of Psa-v expression is mediated by the binding of TEMP-1 to its recognition sites.
The enhancement of GUS activity by the presence of extra TEMP-1 binding sequences in the seed-specific region of *Psa-v*, however, was not strictly additive (Figure 4.8). This suggests that enhancement of activity, as for elements in light-responsive (Gilmartin and Chua, 1990), hormone-responsive (Rogers and Rogers, 1992) and vascular-specific (Keller and Baumgartner, 1991) genes, may be moderated by their position relative to the transcriptional site.

The GUS activities of pea cotyledons transfected by constructs **R-88/-150** and **I-88/-194**, were not as high as those transfected by **I-88/-150**, although they contained as many, or more, regions associated with TEMP-1 binding. The extra TEMP-1 sites in **R-88/-150** and **I-88/-194** are located further upstream to those in **I-88/-150**, suggesting that close spacing of the elements, to each other, or to the transcription start site may be critical. This conclusion is supported by the mutation of individual TEMP-1 binding regions which showed that mutation of the proximal TEMP-1 site had the greatest effect, reducing expression to 22%, whilst mutation of the farthest TEMP-1 site had the lowest effect, reducing expression to 56% (Figure 4.7).

Altering the orientation of regulatory elements to each other can also affect gene expression (Rogers and Rogers, 1992). The TEMP-1 binding regions appear to be functional when they are in the reverse orientation, since region -194/-150, which was defined as a minor positive region by internal mutation (Figure 4.7), also contains a TEMP-1 binding site, but the consensus sequence is in the reverse orientation. Despite this, the element between -194/-150 did not appear to activate transcription as much as the proximal elements, and so the inversion of the orientation of the TEMP-1 sites may reduce the level of enhancement of gene expression. Ligation of the inverted region -88/-306 to **Δ-73-GUS**, however, resulted in equivalent or higher GUS activity than from the ligation of the region in the normal orientation (**R-88/-73**, Figure 4.8). Some of the differences in enhancement might also be moderated by the relative position of the binding sites around the helical structure of the DNA, or by the length of the DNA between binding sites, particularly if DNA-looping or bending is required for the interaction of bound nuclear protein. More detailed analysis will be required to sort out the moderating effects of the position and orientation of these elements.
4.4.4 The role of CATGCATG motifs in seed-specific expression

Deletion of -117/-90 showed that this region contains a strong positive element (Figure 4.5). This sequence can be divided into two subsections; it contains a CACA element and other DNA sequences of the 5' region of vicilin box I, located between -117 and -99, which are recognised by the nuclear protein TEMP-1 (Chapter 3), and a CATGCATG motif between -96 and -89. Mutation of the TEMP-1 site (Δ-129/-96) had a severe effect upon gene expression, reducing it by 78%, but mutation of the tATGCATG sequence (Δ-98/-79) also reduced expression, (to 59%) indicating that this motif may also have a quantitative effect on gene expression.

CATGCATG motifs were initially identified by Dickinson et al. (1988) and are present, with up to one mismatch, in about 90% of the 5' regions of a range of genes that are expressed in the seeds of legumes (Table 1.2). Deletion analyses have shown that the sequences 5' of CATGCATG elements are not sufficient for high-level gene activity (Rerie, 1989; Shirsat et al., 1990; Rerie et al., 1991; Baumlein et al., 1991; Lelievre et al., 1992; Chapter 2; Figure 4.5). Despite this, internal deletion of CATGCATG motifs and shorter CATGCAT/c motifs within the 5' regions of legume storage genes has shown that they may have two roles; to promote high-level expression in seeds (Baumlein et al., 1992; Fujiwara and Beachy, 1992; Chamberland et al., 1992; Lelievre et al., 1992), and also to repress (low-level) expression in leaves (Baumlein et al., 1992; Fujiwara and Beachy, 1992).

The 5' deletion data show that the tATGCATG motif between -117 and -90 has a similar function. Deletion of this region reduced GUS expression from 44% to 20% in cotyledons, and conferred about a 4-fold increase in expression in petals, leaves and stems. Deletion between -73/-61 reduced expression in cotyledons to background levels. This 12 bp sequence contains a second CATGCATG-like sequence, ggTGCATG, which suggests that even shorter CATGCATG motifs may promote expression in seeds. The presence of this second motif, at -73, may also explain why the internal mutations, including ΔH-98/-79, which contained the tATGCATG motif, failed to confer an increase in expression in non-seed organs. It would now be interesting to determine the effect of the mutation of both CATGCATG motifs on expression in seeds and other organs.
A CATGCATG-like motif is also present in a negative element of an oat phytochrome gene. This element is conserved in other phytochrome genes, with the consensus cATGGGcgcncG, and represses gene expression in response to red light (Bruce et al., 1991). A vascular-specific gene encoding bean GRP 1.8, also contains a negative element, the first 7 bp of which are the CATGCATG-like sequence TGCATGC. This 19 bp element controls tissue-specific expression by repressing the expression of element SE1, which otherwise confers expression within non-vascular tissue in stems (Keller and Baumgartner, 1991). Since alternating purine/pyrimidine (R/Y) have the potential to form Z-DNA, it has been proposed that CATGCATG-like elements may regulate expression in many genes (An, 1987; Nathan et al., 1988), including seed-specific genes (Riggs et al., 1989; Aryan et al., 1991), through conformational changes in the DNA structure that might mask or repress DNA-protein binding to adjacent sites. Thus, the CATGCATG motif appears to belong to a larger family of RY-containing negative elements.

4.4.5 Vicilin box I (octanucleotide and CACA) and box II sequences are located within positive regulatory elements.

VICILIN BOX I

The region between -129/-96 of Psa-v contains 30/33 bp of the conserved vicilin box I that is present, in approximately the same location, in most genes encoding 7S proteins (Rerie et al., 1992; Figure 1.4), and has been shown to contain sequences required for binding by TEMP-1. The decrease in expression when this region was mutated, reducing expression to 22% of wild-type activity, indicated that this conserved sequence does indeed have a role in high-level seed-specific gene expression.

Vicilin box I contains two smaller motifs, both of which are present in the less conserved parts of the TEMP-1 binding consensus sequence; an octanucleotide sequence, GCCACt/cTC (Rerie et al., 1992) and a CACA element (Goldberg, 1986; and Figure 1.4). The octanucleotide box is present in almost all legume genes encoding globulins (7S and 11S), and has not been observed within genes that are not expressed in seeds (Table 1.2). The CACA element is a semi-conserved DNA-sequence that is present within many genes that are expressed in seeds and is present at lower frequency in the promoters of many other types of plant genes (Table 1.2). High-level expression in seeds has been correlated with DNA fragments containing of the octanucleotide motif.
(Voelker et al., 1987; Sugaya and Uchimiya, 1992), or the CACA box (Aryan et al., 1991), or both motifs (Chen et al., 1988; Burow et al., 1992). The conservation of vicilin box I and the position of the smaller motifs within it, as well as the fact that vicilin box I seems to be restricted to genes encoding 7S proteins (Table 1.2), suggests that vicilin box I may confer certain gene-specific aspects of regulation. For example, since vicilins are the earliest seed storage protein genes to be expressed (see Figure 5, Rerie et al., 1992), vicilin box I might contribute to a vicilin-specific temporal pattern of expression.

**VICILIN BOX II**

Vicilin box II was also shown to contribute to transient expression of the *Psa-v-GUS* genes in pea seeds. Mutation, between -186/-150, of this 21 bp box (plus 14 adjacent bp) reduced expression in bombarded pea cotyledons by 44%. The vicilin box II sequence has been identified within several vicilin genes (Newbiggin, 1988; Table 1.2) and contains a TEMP-1 binding site (Chapter 3). As shown for *Psa-v* (Figure 4.5), 5' deletion of vicilin box II in the *Gma-α'* gene, between -348/-327, had no significant effect on the detectable level of gene expression (Chen et al., 1986). Similar analysis of a promoter of a gene encoding a heatshock protein, also indicated that two AT-rich regions recognised by a DNA-binding protein did not contribute to gene expression in the context of the 5' region of the heatshock-responsive gene (Czarnecka et al., 1990). But when copies of these binding sites were placed 5' to a truncated (-140) maize *Adh1* promoter they enhanced expression in response to anaerobic induction (Czarnecka et al., 1992). Thus, weak positive elements may not be identifiable with some types of promoter analyses. Consequently, the effectiveness of a positive element may depend upon the gene context, and like vicilin box II and CATGCATG elements, may require interaction with other upstream elements.

The *Pvu-β* gene (which codes for another 7S protein) lacks a vicilin box II sequence, but contains three AG-rich regions (one in inverted orientation) that are similar to the TEMP-1 consensus within vicilin box II (Chapter 3). These AG-1 regions also bind nuclear proteins (Kawagoe and Murai, 1992). This suggests that divergent vicilin box II sequences may be present in other seed storage protein genes, and that these motifs may contribute to gene expression via interaction with specific nuclear proteins, such as TEMP-1 and AG-1.
OTHER ELEMENTS

Not surprisingly, the CAAT and TATA boxes were associated with positive elements. Mutation and deletion of regions containing these sequences, between -88/-73 and -61/-4 reduced gene expression but did not affect tissue specificity (Figures 4.5, 4.6, 4.7, summarised Figure 4.9). TEMP-2 and TEMP-3 bound within the -61/-4 region and so loss of gene expression may have been due to mutation of either or both of these binding sites, and/or the TATA box.

4.4.6 Negative and temporal regulation of Psa-v.

A GENERAL NEGATIVE ELEMENT IS LOCATED BETWEEN -230/-194

Internal mutations in the Psa-v-GUS gene constructs showed, in transient expression assays, that the region between -230/-194 contains a negative element. Mutation of the negative region increased expression in pea cotyledons (Figure 4.7). Expression in leaves bombarded with S\textsuperscript{C}-230/-194 was not quantified. Since, however, deletion between -306 to -194 also increased expression in leaves, petals and stems (Figure 4.6), this suggests that the element between -230/-194 acts as a general negative element, in all tissues (Figure 4.10).

Preliminary results indicated that the deletion Δ-230/-194 enhanced expression in immature cotyledons (10-11 DAF), which might suggest that this deletion affected temporal regulation. Expression of Δ-230/-194, however, appeared to be enhanced even further in mid-maturation cotyledons (data not shown). Thus, the expression in young seeds was probably due to a general increase in expression, rather than the deletion of a negative temporal element.

The phaseolin gene Pvu-β, contains a variety of negative elements, that affect organ-specific, tissue-specific and temporal regulation. These elements have been defined by Bustos et al. (1991) and Burow et al. (1992) in two papers which have slightly conflicting data (see Table 1.1). The negative elements that moderated expression in seeds, located within -518/-418 and -391/-295 (Bustos et al., 1991), and -422/-295 (Burow et al., 1992), did not affect expression in non-seed organs (Burow et al., 1992), but did inhibit expression in immature seeds. Deletion of either element advanced the onset of expression by three days (Bustos et al., 1991). Deletion of the other negative element that has been defined, between -227/-106, resulted in approximately a 10-fold
increase in the low-level of expression of phaseolin in stems and roots, without affecting expression in seeds or leaves (Burow et al., 1992). Furthermore, deletion of a region containing this negative stem/root element resulted in reduction of the concentration of phaseolin mRNA during the mid-late stages of seed development, when the level of this mRNA is usually at a peak (Burow et al., 1992). Hence the negative element between -230/-194 of \textit{Psa-v} does not appear to resemble either type of negative element that has been identified within \textit{Pvu-\beta}. The pea element appeared to act as a general negative element, reducing expression in many organs and different stages of development.

The negative regulatory region of \textit{Psa-v}, between -230/-194, is AT-rich. A light-regulated \textit{Cab-E} gene from \textit{Nicotiniana plumbaginifolia} contains a negative element also located in AT-rich sequences, between -1182/-747 (Castresana et al., 1988). The inhibitory effect of this element may be mediated by the three AT-1 binding sites within this region, which bind to the consensus sequence A/gA/tATTTTTAT/aT/c/a (Datta and Cashmore, 1989). Protein binding sites have been identified in many negative elements, both in genes that are expressed highly in seeds (Riggs et al., 1989; Kawagoe and Murai, 1992) and in other plant genes (Castresana et al., 1988; Datta and Cashmore, 1989; Harrison et al., 1991; Bruce et al., 1991; Cushman and Bohnert, 1992). Seed nuclear proteins were not observed binding to the negative element in the 5' region of \textit{Psa-v}, although leaf nuclear protein, L1, binds to sequences in this element (Chapter 3). This suggests that negative regulation of expression by this element may use a different mechanism to that ascribed to other elements in both light-responsive and other seed-specific genes.

\textbf{TEMP-1 REGIONS MAY MODERATE TEMPORAL REGULATION}

Although the negative element in the \textit{Psa-v} promoter did not appear to affect temporal regulation, preliminary results suggest that the TEMP-1 regions may act to induce expression in immature cotyledons, during the early stages of seed development. Addition of extra TEMP-1 binding sites appeared to increase expression to very high levels in very immature cotyledons (9-10 DAF), as well as increasing activity in older seeds (11-16 DAF). Intense GUS activity was observed, using histochemical and fluorometric assays, within 9-10 DAF cotyledons transfected with each construct containing 5 or 6 TEMP-1 sites (constructs I-88/-150, I-88/-194, R-88/-150, Figure 4.9 and data not shown). Hence the additional TEMP-1 binding sites may affect the temporal expression of the gene, by activating transcription at early stages of seed development, perhaps by overcoming inhibition of \textit{Psa-v} expression in immature seeds.
The $Psa\text{-}v$ TEMP-1 elements appear to have an opposite effect to the temporal elements between -518/-418 and -391/-295 of $Pvu\text{-}\beta$; which repress activity in seeds, particularly in younger tissues (Bustos et al., 1991). Neither do they appear to resemble other temporal elements within $Pvu\text{-}\beta$ (-295/-106, Burow et al., 1992) or $Gma\text{-}\alpha'$ (-245/239, Fujiwara and Beachy, 1992), which confer expression during mid to late maturation, without affecting low-level early expression.

Since the additional TEMP-1 regions may affect temporal expression, and the level of expression conferred by TEMP-1 sites appears to be moderated by position and/or orientation of TEMP-1 elements, this suggests that different seed-storage proteins genes may have developed different temporal expression programs, through the duplication and change in the position and orientation of a small variety of conserved DNA motifs within the 5' regions of these genes. Further study of the effect of extra TEMP-1 binding sites would be best investigated in stably transformed pea plants, to determine when, and in which cell types and organs, enhanced gene expression occurs.

4.4.7 Summary

The location of the principal regulatory elements that were identified within the seed-specific region of $Psa\text{-}v$ correlated with the position of a number of conserved motifs, including CATGCATG, vicilin box I and the ubiquitous CAAT and TATA boxes. Vicilin box II was shown to contain a minor positive element. The results demonstrate that these conserved sequences have a regulatory role in gene function; an idea that was initially proposed when seed-specific sequences were first identified (Gatehouse et al., 1986; Goldberg et al., 1986, Dickinson, 1988; Newbigin, 1988).

DNA-binding studies (Chapter 3) had shown that the temporally regulated and seed-specific nuclear proteins, TEMP-1, TEMP-2 and TEMP-3 bound to sequences within the seed-specific region. The correlation of the principal regulatory elements, with the regions that contained binding sites for these proteins supports the proposal that these nuclear proteins may have a role in the regulation of $Psa\text{-}v$ gene expression.
CHAPTER 5. CONCLUSIONS

The aims of this project were to investigate how the \( Psa-v \) gene encoding the vicilin seed storage protein is regulated; to determine what DNA sequences are important for the seed-specific, temporal and quantitative aspects of the expression of this gene, to locate specific recognition sites for nuclear proteins. The activity of such DNA-protein binding factors during seed development and in other organs was also examined, to gain an understanding of how these putative \textit{trans}-acting factors themselves may be regulated.

It had been shown earlier that gene constructs containing the intact 5' region of \( Psa-v \), and the 5' regions of other seed storage protein genes, were expressed specifically in the seed of transgenic plants (see Goldberg \textit{et al.}, 1989; and Chapter 1, for review). Thus, genes that are highly expressed in seeds must contain conserved \textit{cis}-acting sequences that can be recognised and acted upon by transcription factors from different plant species.

Seed-specific and gene-specific DNA sequences, that might contain \textit{cis}-acting elements which regulate the expression of these genes, have been identified in 5' regions of seed storage protein genes. Initial promoter analyses, however, indicated that these conserved sequences were not sufficient to confer detectable levels of expression of \( Psa-v \) (Newbigin, 1988) and other legume seed storage protein genes (Shirsat \textit{et al.}, 1989; Shirsat \textit{et al.}, 1990; Aerie \textit{et al.}, 1991). Thus, either these conserved DNA motifs have no regulatory role in gene expression, or else interaction(s) with additional DNA element(s) is required.

5.1 The 5' region of \( Psa-v \) contains many \textit{cis}-acting elements and DNA-protein binding sites that are located in conserved or repeated DNA sequences.

It was shown in Chapter 2, by analysis of \( Psa-v \) gene constructs in stably transformed tobacco plants, that the conserved DNA sequences vicilin box I and vicilin box II and two CATGCATG sequences (tATGCATG and ggTGCATG) were present in a region of DNA that conferred both seed-specific and temporally regulated expression. Since further deletion of a region between -306 to -97, containing the vicilin box conserved sequences, reduced expression below detectable levels, this suggested that these two sequences may indeed contribute to gene expression, but queried the role of the CATGCATG sequences in this gene. Two positive quantitative regions were
located upstream of position -306. Thus, additional DNA elements were required for full expression of the Psa-v gene.

DNA-protein binding assays, using fragments from the quantitative and seed-specific/temporal elements, located DNA sequences that were recognised and bound by nuclear proteins from seeds (Chapter 3). The region between -773 and +1 of Psa-v contained at least seven regions that were bound by nuclear proteins.

Both the major (-773/-425) and the minor (-424/-307) quantitative elements contain a copy of an AT-rich DNA repeat. Sequences in each AT-rich repeat containing the consensus sequence GATAa/gATAAT were bound by a nuclear protein termed AT-2. Short AT-rich sequences that bind nuclear proteins have been shown to act as cis-acting elements in genes expressed in seeds and in other organs (Bustos et al., 1989; Czarnecka et al., 1992; see Chapter 1 for review). Thus, the occurrence of an AT-2 binding site in each quantitative element suggests that the binding of AT-2 to each site may enhance transcription. Experiments that test the effect of mutagenesis and duplication of the AT-2 sites are required to determine whether these are the cis-acting elements within the quantitative regions and whether the presence of extra AT-2 binding sites has an additive, a geometric (synergistic) or no effect on gene expression.

Three regions, -186/-150, -153/-117, -129/-96, that encompassed vicilin box II and the octanucleotide and CACA regions of vicilin box I, respectively, were shown to be required for binding of the nuclear protein TEMP-1. It is proposed that TEMP-1 binds to DNA sequences in the consensus TTtttTtcAtTTTgacacat, which is present in each of the three TEMP-1 binding regions (Chapter 3). The transient expression of three types of Psa-v-GUS gene constructs, (with deletion, mutation and duplication of the TEMP-1 binding regions), was analysed in pea cotyledons that had been transfected using particle bombardment (Chapter 4). This showed that the TEMP-1/vicilin box regions contained cis-acting elements that were essential for high-level gene expression in pea seeds. The level of enhancement of gene expression that was conferred by additional copies of the TEMP-1 binding regions differed among constructs. The orientation and position of the TEMP-1 binding sites, and possibly the distances between individual TEMP-1 binding regions, may all be factors that moderate the effectiveness of these elements.
Since the DNA sequences that were required for recognition of the TEMP-1 nuclear protein were found in the vicilin box I and II sequences, this also raises the question as to whether many of the cis-acting sequences that have been conserved within Psa-ν and other seed-specific genes are actually part of different TEMP-1 binding sites.

Binding sites for nuclear proteins were also present between -61 to -4 of the Psa-ν gene (Chapter 3). Although this region was not sufficient for high-level expression (Chapters 2 and 4), mutation of this sequence reduced expression in seeds by 79%, indicating that the DNA-protein binding sites, and/or other sequences in this region were important for expression. It is possible that, like the nuclear factor binding to the TATA box of the Pvu-β gene, (Kawagoe and Murai, 1992), one or both of the TEMP-1 and TEMP-2 proteins may bind to the Psa-ν TATA box.

5.2 The binding activity of TEMP-1, TEMP-2 and TEMP-3 correlates with activation of Psa-ν gene expression.

When the binding activity of nuclear proteins from seeds at different stages of development was examined (Chapter 3), it was shown that the peaks in intensity of TEMP-1, TEMP-2 and TEMP-3 binding, between 11 to 15 DAF, correlated with the initial increases in steady-state levels of vicilin mRNA. This indicated that these proteins may have a role in the temporal regulation of Psa-ν gene expression. AT-2 binding activity, however, was detected in extracts from seeds at all developmental stages, except from the youngest seed. This supports the proposition that the quantitative elements, which contain the AT-2 binding sites, only enhance gene expression, rather than moderate temporal or seed-specific expression.

The nuclear protein AT-2, like the binding proteins TEMP-1, TEMP-2 and TEMP-3 was not detected in leaf nuclear extract. This suggested that the nuclear proteins may be specific to seeds. Regulation of transcriptional activity by the regulation of the abundance of the required transcription factors, appears to be common to many genes (Latchman, 1990; Falvey and Schibler, 1991). It appears, however, that activation of the Psa-ν gene may involve two other forms of transcriptional regulation. One is the specific inhibition of the DNA-binding activity of TEMP-1 and AT-2 in leaves, the second may involve inhibition and activation mediated by the CATGCATG elements, perhaps by induced changes in the conformation of the DNA.
5.3 Activation of *Psa-v* expression in leaves may be regulated by a protein that inhibits DNA-protein binding of the "seed-specific" binding proteins.

It was shown (Chapter 3), by mixing nuclear extracts from different sources, that the leaf nuclear extract contained an inhibitor that specifically inhibited binding of AT-2, TEMP-1 to *Psa-v* gene fragments. This factor also inhibited the binding of nuclear protein to a quantitative region of another pea seed storage protein gene, *LegA1*, (Chapter 3; Appendix II), but had little or no effect on binding of the nuclear proteins OCSTF and EBNA-1 to their target sites. Hence the inhibition of TEMP-1 and AT-2 binding activity to *Psa-v* may be an important mechanism in repressing the activation of the gene in leaves.

The specific inhibitory protein present in leaves may binds to the "seed-specific" nuclear proteins, masking the DNA binding site of these seed nuclear proteins. Examples of DNA-binding proteins that are regulated in this fashion include MyoD1 (Davis et al., 1987; Benezra et al., 1990) and other transcription factors (Foulkes et al., 1991; Nakabeppu and Nathans, 1991; Foulkes et al., 1992).

There may be little homology between the abundant AT-rich binding proteins and the less abundant TEMP-1 binding proteins. Thus, it may be more likely that the inhibitor is an enzyme that modifies the seed binding protein at the DNA-binding domain. If the inhibition of the DNA-binding activity of TEMP-1 and AT-2 could be reversed, then this could also be used to establish whether the absence of binding activity in the leaf extract is due solely to inhibition, or whether the abundance of the TEMP-1 and AT-2 proteins are also reduced.

The binding activity of AT-1 (Datta and Cashmore, 1989) 3AF3 (Sarokin and Chua, 1992) and GBF1 (Klimczak et al., 1992) that bind to sequences in light-responsive genes and many other eukaryotic DNA-binding proteins (Hunter and Karin, 1992) are regulated by their phosphorylation status. The addition of phosphotases and protein kinases to leaf and seed extracts may help to resolve whether the binding of the seed nuclear proteins is regulated by phosphorylation. Further experiments are also required to ascertain whether the leaf inhibitor is present in other vegetative organs, and whether it affects the binding of TEMP-2 and TEMP-3. The low level of binding in extracts from immature seeds suggests that peas at early stages of development may contain (steadily decreasing) levels of inhibitor.
5.4 How do the CATGCATG sequences affect gene regulation?

Not all the cis-acting elements that were identified were shown to contain DNA-protein binding sites. These included a minor negative element between -231/-194, and positive elements between -98/-79, -88/-73 and -73/-61. Two of these regions contained CATGCATG elements; tATGCATG and ggTGCATG, located at -96 and -68, respectively. CATGCATG elements are located in the 5' regions of most genes that are highly expressed in seeds (Dickinson et al., 1988; and Table 1.2), but it is yet to be shown that this sequence is bound specifically by nuclear proteins (Jofuku et al., 1987; Allen et al., 1989; Bustos et al., 1989; Jofuku et al., 1987; Allen et al., 1989; Bustos et al., 1989; Jofuku et al., 1987; Allen et al., 1989; Bustos et al., 1989; Jofuku et al., 1987; Allen et al., 1989; Bustos et al., 1989; Jofuku et al., 1987; Allen et al., 1989; Lessard et al., 1991; Meakin and Gatehouse, 1991; Fujiwara and Beachy, 1992; Kawagoe and Muraiz, 1992; Chapter 3).

It is possible that a CATGCATG-binding protein, if it exists, is not at all abundant, or is highly labile, and so is unlikely to be detected. Alternatively, it has been proposed (Riggs et al., 1989; Aryan et al., 1991), based upon the ability of DNA with alternating purine-pyrimidine nucleotides, such as CATGCATG, to form Z-DNA, that the CATGCATG element may act via conformational changes in the DNA, rather than by interaction with a specific transcription factor.

The CATGCATG elements in the Psa-v gene flank the CAAT and TATA boxes, both of which (not surprisingly) were shown to be located in DNA sequences that were required for high-level gene activity (Chapter 4). These results lend themselves to a model (see Figure 5.1) whereby the Z-DNA conformation of the CATGCATG elements masks the TATA and CAAT boxes preventing binding of factors and transcriptional initiation. This model, by changes in Z-DNA conformation might also explain why AT-2 binding activity is present when vicilin mRNA is low or absent. For example, AT-2 might continue to bind to Psa-v, during later stages of seed development, but have no effect upon transcription because of the absence of the transcription complex. It is proposed in the model that the inhibitory Z-DNA conformation is altered to B-DNA by the binding of one or more of the TEMP proteins, which bind near to the CATGCATG sites. Thus the presence of the TEMP proteins would control the expression of the gene during the development of the seed.
Figure 5.1
Proposed model of the transcriptional regulation of \textit{Psa-v}.

This model is based upon several assumptions:
- That the concentration of the inhibitor factor is high in all non-seed tissues, and decreases during early seed development.
- That the AT-2, TEMP-1, -2 and -3 DNA-binding proteins are either inactivated by the inhibitor at these stages \textbf{and/or} are not present in high concentration.
- That the CATGCATG motifs can adopt a Z-DNA conformation, preventing access to the CAAT and TATA boxes. Under these circumstances DNA-proteins are unable to bind and/or transactivate expression.

DNA elements and protein binding sites in the region -773 to +1 of \textit{Psa-v} are as described in Figures 3.1 and 4.10. Binding proteins: AT-2, TEMP (1, 2, 3) binding inhibitor (I), (putative) CAAT box binding protein (C), and the transcription complex containing RNA polymerase II (pol II), are indicated. The CATGCATG boxes are shown in "Z-DNA" conformation as zigzags and in "B-DNA" conformation as triangles.

\textbf{Leaf and immature seed}
Transcription in the vegetative organs is prevented (X) by the presence of the binding inhibitor protein \textbf{and/or} a reduced concentration of the DNA binding proteins. The Z-DNA conformation of the CATGCATG elements also prevents binding of TFIID and other general proteins associated with RNA polymerase.

\textbf{Mid-mature seed}
As the seed develops, the concentration of the inhibitor diminishes, and/or transcription of the DNA-binding proteins is stimulated, resulting in high concentrations of active binding factors.

Binding of a specific factor, for example TEMP-1, (or some other mechanism), changes the conformation of the CATGCATG motifs to B-DNA, permitting access to the proximal promoter elements. AT-2, TEMP-2 and TEMP-3 also bind. Subsequent interactions between the binding proteins permit binding of RNA polymerase II and other proteins of the transcription initiation complex, resulting in transcription of \textit{Psa-v} (•).

\textbf{Mature seed}
As the seeds develop further, the abundance of TEMP-1 and TEMP-2 reduces dramatically. The reduction in binding of these TEMP-1 causes the CATGCATG motifs to adopt Z-DNA conformation again, and so, despite the presence of AT-2 and (to some extent) TEMP-3 binding proteins, transcription is inhibited.
This page contains diagrams illustrating the transcriptional regulation of the Tested gene in different stages of seed development. The diagrams show the interaction of different transcription factors (AT-2 and pol II) with the tested gene promoter in leaf and immature seed, mid-mature seed, and mature seed stages.

1. **Leaf and Immature Seed**: The diagram illustrates the interaction of transcription factors AT-2 and pol II with the Tested gene promoter. The factors are shown to bind to specific sites (1, 2, 3) on the promoter, indicating overlapping positive regulatory regions.

2. **Mid-Mature Seed**: The diagram shows the interaction of AT-2 and pol II with the Tested gene promoter. The factors are depicted as interacting with the promoter in a manner that is consistent with positive regulation.

3. **Mature Seed**: The diagram indicates the interaction of AT-2 and pol II with the Tested gene promoter. The factors are shown to bind to the promoter in a way that is consistent with positive regulation, with overlapping positive regulatory regions.

These diagrams are part of a study on the transcriptional regulation of the Tested gene in seed development, highlighting the role of specific transcription factors in the regulation of gene expression at different developmental stages.
Deletion of the region of $Psa-v$ between -117 to -90, which contains the sequence tATGCATG, resulted in a small increase in expression in transiently transformed petals and leaves and a significant decrease in expression in seeds. Further deletion of 12 bp which included the sequence ggTGCATG at -68, reduced expression to background levels. Internal mutation of an 18 bp sequence containing tATGCATG also significantly reduced gene expression, by 41%. Thus the presence of the CATGCATG motifs in $Psa-v$ correlated with high gene expression in seeds (although less than the TEMP-1/vicilin box and TEMP-2/-3 binding sites) and repression of low-level activity in non-seed organs.

These results support other data that have shown that mutation or internal deletion of CATGCATG elements results in significant loss of activity of seed-specific gene expression (Baumlein et al., 1992; Chamberland et al., 1992; Fujiwara and Beachy, 1992; Lelievre et al., 1992) and can also increase the level of activity in the leaves (Baumlein et al., 1992; Fujiwara and Beachy, 1992).

Since the mutation of the tATGCATG motif in the region -306/+39 of $Psa-v$ did not result in a visible increase in expression in non-seed organs, either the presence of the other CATGCATG motif in this context was able to inhibit expression, or other regulatory mechanisms (as discussed above) may prevent transcription in non-seed tissues. These questions would be best tested by analysing the expression of gene constructs with mutations at both CATGCATG sites, preferably in stably transformed plants, where weak expression in vegetative organs could be more readily detected and quantified.

### 5.5 Applications

An objective of the analysis of the expression of seed storage protein genes is to garner information that will aid in achieving high-level expression of foreign genes (encoding proteins that will give added disease or insect resistance, or that will improve the nutritional quality) in the seeds of transgenic plants. Analysis of the transient activity of different $Psa-v$-GUS gene constructs indicates that a 5' region that would increase expression of target genes in transgenic seeds, perhaps by 10 fold or more, could be made by the duplication of some of the cis-acting elements, especially the TEMP-1/vicilin box I region between -150 to -88, and also the AT-2 binding sites. Deletion of the weak negative element could also prove useful.
To understand how seed-storage protein genes are regulated, by different cell-specific and temporally-regulated factors, it is necessary, once the principal cis-acting elements are localised, to identify and investigate trans-acting factors that bind to these elements - to discover what regulates the regulators. It will be intriguing to find out more about how the inhibitory factor in leaves affects DNA-binding activity, how the activity of the TEMP proteins during seed development is regulated, and also, to determine the mechanism by which the CATGCATG elements work.
Appendix I. Frequency of putative seed-specific motifs in plant promoters

Appendix to Table 1.2. Number of seed-specific and other motifs within the 5' region of various plant genes. Only sequences containing 100 bp or more were analysed. The majority of sequences were compiled by Leon Dure, and sources and the gene names for the sequences are indicated, when known.

<table>
<thead>
<tr>
<th>LEGUMES</th>
<th>Storage</th>
<th>7S</th>
<th>ALL</th>
<th>Legumes</th>
<th>Cereals</th>
<th>Other</th>
<th>REFERENCES</th>
</tr>
</thead>
<tbody>
<tr>
<td>SEED</td>
<td>catgcatg</td>
<td>CACA</td>
<td>ANCCCA</td>
<td>oct</td>
<td>vlc I</td>
<td>vlc II</td>
<td>leg box</td>
</tr>
<tr>
<td>Pea-v</td>
<td>307</td>
<td>5</td>
<td>Pea Victin</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Cva-a</td>
<td>590</td>
<td>5</td>
<td>Pea Convitlin</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Cva-b</td>
<td>1000</td>
<td>5</td>
<td>Pea Convitlin</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Pvu-8</td>
<td>956</td>
<td>5</td>
<td>Phaseolus Viclin</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Gma-a'</td>
<td>973</td>
<td>5</td>
<td>Soybean Viclin</td>
<td>1</td>
<td>2</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>CG4</td>
<td>153</td>
<td>5</td>
<td>Soybean Viclin</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Vlc-1</td>
<td>418</td>
<td>5</td>
<td>Vicia faba Viclin</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>11S LegA</td>
<td>500</td>
<td>5</td>
<td>Pea Legumin</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>LegA2</td>
<td>1582</td>
<td>5</td>
<td>Pea Legumin</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>LegB</td>
<td>300</td>
<td>5</td>
<td>Pea Legumin</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>LegC</td>
<td>300</td>
<td>5</td>
<td>Pea Legumin</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Le1</td>
<td>617</td>
<td>5</td>
<td>Pea Legumin</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SLV</td>
<td>691</td>
<td>5</td>
<td>Soybean Glycin</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Gyl</td>
<td>638</td>
<td>5</td>
<td>Soybean Glycin</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Gy2</td>
<td>801</td>
<td>5</td>
<td>Soybean Glycin</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Gy3</td>
<td>297</td>
<td>5</td>
<td>Soybean Glycin</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Gy4</td>
<td>300</td>
<td>5</td>
<td>Soybean Glycin</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Leb4</td>
<td>735</td>
<td>5</td>
<td>Vicia Legumin</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2S</td>
<td>220</td>
<td>5</td>
<td>Pea Albumin</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Lectins</td>
<td>lecA</td>
<td>476</td>
<td>5</td>
<td>Pea Lectin</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>ddic1</td>
<td>374</td>
<td>5</td>
<td>Phaseolus Lectin</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>ddic2</td>
<td>528</td>
<td>5</td>
<td>Phaseolus Lectin</td>
<td>6</td>
<td>1</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>nSTP-2</td>
<td>264</td>
<td>5</td>
<td>Phaseolus Lectin</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Pddec1</td>
<td>500</td>
<td>5</td>
<td>Phaseolus Lectin</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Pddec2</td>
<td>500</td>
<td>5</td>
<td>Phaseolus Lectin</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Le1</td>
<td>968</td>
<td>5</td>
<td>Soybean Lectin</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Other</td>
<td>Aux28</td>
<td>484</td>
<td>5</td>
<td>Soybean Auxin Regulated</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Aux22</td>
<td>683</td>
<td>5</td>
<td>Soybean Auxin Regulated</td>
<td>5</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>BB13.10</td>
<td>142</td>
<td>5</td>
<td>Soybean B.B. Inhibitor</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>NONSEED Histonenes</td>
<td>LEGUMES</td>
<td>CEREALS</td>
<td>OTHER REFERENCES</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------------------</td>
<td>---------</td>
<td>---------</td>
<td>------------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Heatshock</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GmHesp17.6</td>
<td>419</td>
<td>5'</td>
<td>Soybean 17.6 Heat Shock</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>GmHesp17.5E</td>
<td>594</td>
<td>5'</td>
<td>Soybean 17.5 Heat Shock</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>HS176</td>
<td>1070</td>
<td>5'</td>
<td>Soybean 17.6 Heat Shock</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>GmHesp18.5C</td>
<td>1920</td>
<td>5'</td>
<td>Soybean 18.5 Heat Shock</td>
<td>1</td>
<td>5</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>ShS200</td>
<td>1512</td>
<td>5'</td>
<td>Soybean 22KD Heat Shock</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ShS27</td>
<td>658</td>
<td>5'</td>
<td>Soybean 27KD Heat Shock</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>GmHesp17.3B</td>
<td>594</td>
<td>5'</td>
<td>Soybean 17.3 Heat Shock</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>SSOYHSP</td>
<td>963</td>
<td>5'</td>
<td>Soybean 17.5 Heat Shock</td>
<td>0</td>
<td>3</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Other</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SGA19</td>
<td>740</td>
<td>5'</td>
<td>Alfalfa Glutamine Synth.</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>lsl</td>
<td>191</td>
<td>5'</td>
<td>Soybean Leghemoglobin</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>lba</td>
<td>177</td>
<td>5'</td>
<td>Soybean Leghemoglobin</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5LEG8</td>
<td>178</td>
<td>5'</td>
<td>Soybean Leghemoglobin</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>lbc</td>
<td>117</td>
<td>5'</td>
<td>Soybean Leghemoglobin</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>lbc3</td>
<td>769</td>
<td>5'</td>
<td>Soybean Leghemoglobin</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SLURG</td>
<td>137</td>
<td>5'</td>
<td>Lupin 5.8s rRNA</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>nodulin-23</td>
<td>799</td>
<td>5'</td>
<td>Soybean Nodulin 23</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>nodulin-20</td>
<td>557</td>
<td>5'</td>
<td>Soybean Nodulin 20</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>nodulin-22</td>
<td>521</td>
<td>5'</td>
<td>Soybean Nodulin 22</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>nod35</td>
<td>311</td>
<td>5'</td>
<td>Soybean Nodulin 35</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SPEADH</td>
<td>788</td>
<td>5'</td>
<td>Pea ADH1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>ABB6</td>
<td>183</td>
<td>5'</td>
<td>Pea Chloro. A/B</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5PGS</td>
<td>120</td>
<td>5'</td>
<td>Phaseolus Glutamine Synth.</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5PGSB</td>
<td>116</td>
<td>5'</td>
<td>Phaseolus Glutamine Synth.</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Lbc2</td>
<td>123</td>
<td>5'</td>
<td>Phaseolus Leghemoglobin</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>rbc3-E9</td>
<td>1081</td>
<td>5'</td>
<td>Pea Small Subunit</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>rbc5-3A</td>
<td>430</td>
<td>5'</td>
<td>Pea Small Subunit</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>rbc3-S</td>
<td>328</td>
<td>5'</td>
<td>Pea Small Subunit</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SbPRP1</td>
<td>1123</td>
<td>5'</td>
<td>Soybean Proline-Rich</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>SbPRP3</td>
<td>1201</td>
<td>5'</td>
<td>Soybean Proline-Rich</td>
<td>3</td>
<td>0</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>nodulin-24</td>
<td>734</td>
<td>5'</td>
<td>Soybean Nodulin 24</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>rbc34</td>
<td>715</td>
<td>5'</td>
<td>Soybean Small Subunit</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Ss33</td>
<td>265</td>
<td>5'</td>
<td>Soybean Small Subunit</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SAC1</td>
<td>225</td>
<td>5'</td>
<td>Soybean Actin</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>psAAC3</td>
<td>195</td>
<td>5'</td>
<td>Soybean Actin</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>DICOTS</td>
<td>Storage</td>
<td>catg</td>
<td>catgcag</td>
<td>CACA</td>
<td>ANCCCA</td>
<td>oct</td>
<td>vic I</td>
</tr>
<tr>
<td>--------</td>
<td>---------</td>
<td>------</td>
<td>---------</td>
<td>------</td>
<td>--------</td>
<td>-----</td>
<td>-------</td>
</tr>
<tr>
<td>SEED</td>
<td>A72</td>
<td>547</td>
<td>5'</td>
<td>Cotton vicillin</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>B72</td>
<td>540</td>
<td>5'</td>
<td>Cotton Vicillin</td>
<td>1</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>CRC</td>
<td>115</td>
<td>5'</td>
<td>Arabidopsis Legumin</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>CRA1</td>
<td>196</td>
<td>5'</td>
<td>Arabidopsis Legumin</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>CRA2</td>
<td>204</td>
<td>5'</td>
<td>Arabidopsis Legumin</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>CRB</td>
<td>377</td>
<td>5'</td>
<td>Arabidopsis Legumin</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>REFERENCES</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Dure-EMBL-M19378</td>
<td>1</td>
</tr>
<tr>
<td>Dure-EMBL-M16936</td>
<td>1</td>
</tr>
<tr>
<td>Pang (1989)</td>
<td>1</td>
</tr>
<tr>
<td>Pang (1989)</td>
<td>1</td>
</tr>
<tr>
<td>Pang (1989)</td>
<td>1</td>
</tr>
<tr>
<td>Pang (1989)</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>LECTINS</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>D11</td>
<td>307</td>
</tr>
<tr>
<td>D113</td>
<td>400</td>
</tr>
<tr>
<td>D19</td>
<td>937</td>
</tr>
<tr>
<td>D29</td>
<td>1533</td>
</tr>
<tr>
<td>D34</td>
<td>1228</td>
</tr>
<tr>
<td>D7</td>
<td>537</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>OTHER</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>NONSEED HISTONES</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>SARAb3112</td>
<td>405</td>
</tr>
<tr>
<td>SARAB32</td>
<td>368</td>
</tr>
<tr>
<td>SARAB42</td>
<td>608</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>STORAGE</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>pat21</td>
<td>2312</td>
</tr>
<tr>
<td>pat51</td>
<td>735</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>HEATSHOCK</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A8140</td>
<td>218</td>
</tr>
<tr>
<td>A8165</td>
<td>146</td>
</tr>
<tr>
<td>A8180</td>
<td>207</td>
</tr>
<tr>
<td>chs</td>
<td>1179</td>
</tr>
<tr>
<td>ocdh</td>
<td>1022</td>
</tr>
<tr>
<td>pAVS33</td>
<td>273</td>
</tr>
<tr>
<td>podC5A1</td>
<td>750</td>
</tr>
<tr>
<td>plGAB19/H5c</td>
<td>654</td>
</tr>
<tr>
<td>atp5-2</td>
<td>392</td>
</tr>
<tr>
<td>SGR</td>
<td>593</td>
</tr>
<tr>
<td>SPIN1</td>
<td>799</td>
</tr>
<tr>
<td>LlK</td>
<td>305</td>
</tr>
<tr>
<td>PR-TA</td>
<td>1508</td>
</tr>
<tr>
<td>tbcS-11A</td>
<td>279</td>
</tr>
<tr>
<td>rbcS-8</td>
<td>484</td>
</tr>
<tr>
<td>TP-1</td>
<td>715</td>
</tr>
<tr>
<td>nts23</td>
<td>1043</td>
</tr>
<tr>
<td>CEREALS</td>
<td>5WH7TS</td>
</tr>
<tr>
<td>---------</td>
<td>--------</td>
</tr>
<tr>
<td></td>
<td>5OAIGLO</td>
</tr>
<tr>
<td></td>
<td>alpha-1Y</td>
</tr>
<tr>
<td></td>
<td>qf-1</td>
</tr>
<tr>
<td></td>
<td>pW8233</td>
</tr>
<tr>
<td></td>
<td>L311A</td>
</tr>
<tr>
<td></td>
<td>L311B</td>
</tr>
<tr>
<td></td>
<td>q2195c (299)</td>
</tr>
<tr>
<td></td>
<td>q215A</td>
</tr>
<tr>
<td></td>
<td>ZE19</td>
</tr>
<tr>
<td></td>
<td>Z4</td>
</tr>
<tr>
<td></td>
<td>Z7</td>
</tr>
<tr>
<td></td>
<td>ZA1</td>
</tr>
<tr>
<td></td>
<td>C22A1</td>
</tr>
<tr>
<td></td>
<td>pML1</td>
</tr>
<tr>
<td></td>
<td>pHS216</td>
</tr>
<tr>
<td></td>
<td>Glu-D1-1</td>
</tr>
<tr>
<td></td>
<td>G2</td>
</tr>
<tr>
<td></td>
<td>Glu-D1-1</td>
</tr>
<tr>
<td></td>
<td>pBRH184</td>
</tr>
</tbody>
</table>

| Lectins | waxy | 1232 | 5'   | Malte Waxy Locus | 0   | 0       | 1       | 0     | 0         | 0       | 0       | Klosgen (1986) |
|         | alevuin | 1457 | 5'   | Barley Aleurain | 1   | 2       | 5       | 0     | 0         | 0       | 0       | Whittier (1987) |
|         | 5AMY18 | 1001 | 5'   | Barley Amylase | 1   | 0       | 2       | 0     | 0         | 0       | 0       | Whittier (1988) |
|         | 5AMY18ta | 709 | 5'   | Barley Amylase | 0   | 0       | 2       | 0     | 0         | 0       | 0       | Whittier (1987) |
|         | Amy32b | 689  | 5'   | Barley Alpha-Amylase | 0   | 0       | 2       | 0     | 0         | 0       | 0       | Whittier (1987) |
|         | Amy2/54 | 1000 | 5'   | Barley Amylase | 0   | 1       | 3       | 0     | 0         | 0       | 0       | Whittier (1987) |
|         | 5AMY4a | 468  | 5'   | Barley Amylase | 0   | 1       | 0       | 0     | 0         | 0       | 0       | Whittier (1987) |
|         | alpha-Amy1 | 242 | 5'   | Barley Alpha-Amylase | 0   | 1       | 0       | 0     | 0         | 0       | 0       | Knox (1987) |
|         | alpha-Amy2 | 239 | 5'   | Barley Alpha-Amylase | 0   | 1       | 0       | 0     | 0         | 0       | 0       | Knox (1987) |
|         | 5CARBOX | 1000 | 5'   | Barley Carboxypeptidase | 2   | 0       | 0       | 0     | 0         | 0       | 0       | Bouloubrine, unpub. |
|         | A121 | 830  | 5'   | Barley Cathepsin | 1   | 2       | 1       | 0     | 0         | 0       | 0       | Celjuco (1993) |

| Histones | HMC14 | 350  | 5'   | Malte Histone H4 (c14) | 0   | 0       | 1       | 0     | 0         | 0       | 0       | Chaubet (1986) |
|         | HMC4 | 525  | 5'   | Malte Histone H3 (c4) | 0   | 0       | 2       | 0     | 0         | 0       | 0       | Chaubet (1986) |
|         | HMC7 | 372  | 5'   | Malte Histone H4 (c7) | 0   | 0       | 4       | 0     | 0         | 0       | 0       | Chaubet (1986) |
|         | 5HS | 421  | 5'   | Malte Histone H3 (c2) | 1   | 0       | 3       | 0     | 0         | 0       | 0       | Chaubet (1986) |
|         | RH3-17 | 140  | 5'   | Rice Histone H3 | 0   | 0       | 2       | 0     | 0         | 0       | 0       | Peng (1985) |
|         | Th012 | 247  | 5'   | Wheat Histone H3 | 0   | 0       | 2       | 0     | 0         | 0       | 0       | Tabata (1984) |
|         | Th011 | 735  | 5'   | Wheat Histone H4 | 0   | 0       | 4       | 0     | 0         | 0       | 0       | Tabata (1983) |
### Sequence of motifs

<table>
<thead>
<tr>
<th>Name</th>
<th>Allele Matched</th>
<th>Sequence</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>catgcgtg</td>
<td>1</td>
<td>CAAGCTAG</td>
<td>Dickinson (1988)</td>
</tr>
<tr>
<td>GACA</td>
<td>0</td>
<td>AACACA</td>
<td>Goldberg (1986)</td>
</tr>
<tr>
<td>ADCCCA</td>
<td>0</td>
<td>ANCCCA D = A/C/T</td>
<td>Chen (1986)</td>
</tr>
<tr>
<td>oct</td>
<td>0</td>
<td>GCCAYTyT = C/T</td>
<td>Lycett (1985), Rele (1992)</td>
</tr>
<tr>
<td>vlc I</td>
<td>3</td>
<td>GCCAGN TCAATT/T GTNCATTINAAACACNCGT</td>
<td>Gatehouse (1988), this study</td>
</tr>
<tr>
<td>vlc II</td>
<td>0</td>
<td>NTIC AATA AAT CAAA AAA A</td>
<td>Bown (1988), Newbiggin (1988)</td>
</tr>
<tr>
<td>G-Box</td>
<td>(1)</td>
<td>NACGTG</td>
<td>Giuliano (1988), Gilmore (2000, review)</td>
</tr>
</tbody>
</table>

### CEREALS

<table>
<thead>
<tr>
<th>NONSEED Heatshock</th>
<th>mhsp-70</th>
<th>269</th>
<th>5</th>
<th>Malate 70 KD Heat Shock</th>
<th>catgcgtg</th>
<th>CACA</th>
<th>ANCCCA</th>
<th>oct</th>
<th>vlc I</th>
<th>vlc II</th>
<th>leg box</th>
<th>-300 box</th>
<th>G-box</th>
<th>REFERENCES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heatshock</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Rochester (1986)</td>
</tr>
<tr>
<td>Other</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SWHCAB1</td>
<td>915</td>
<td>5</td>
<td></td>
<td>Wheat Chlo. A/B PtN.</td>
<td>1</td>
<td>4</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>not published</td>
</tr>
<tr>
<td>SWHCAB2</td>
<td>200</td>
<td>5</td>
<td></td>
<td>Wheat Chlo. A/B PtN.</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>Lampia (1985)</td>
</tr>
<tr>
<td>Adh-1</td>
<td>414</td>
<td>5</td>
<td></td>
<td>Malate 124. Dehydro.</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>Dennis (1982)</td>
</tr>
<tr>
<td>Adh-2</td>
<td>414</td>
<td>5</td>
<td></td>
<td>Malate 124. Dehydro.</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>Dennis (1982)</td>
</tr>
<tr>
<td>Maen</td>
<td>221</td>
<td>5</td>
<td></td>
<td>Malate Actin</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>Shah (1983)</td>
</tr>
<tr>
<td>GST-1</td>
<td>245</td>
<td>5</td>
<td></td>
<td>Malate Glutathione Trans.</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>Shah (1983)</td>
</tr>
<tr>
<td>A1</td>
<td>686</td>
<td>5</td>
<td></td>
<td>Malate Anthocyanin Bio.</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>Schwarz-Sommer (1987)</td>
</tr>
<tr>
<td>Mt11</td>
<td>146</td>
<td>5</td>
<td></td>
<td>Malate Trisulf 4 Isom.</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>Marchionni (1986)</td>
</tr>
</tbody>
</table>
APPENDIX II. ANALYSIS OF CONSERVED 5' REGIONS WITHIN VICILIN AND LEGUMIN GENES FROM PEA.

Introduction

Seed storage proteins comprise the major proportion of protein in mature seeds. The expression of these genes is confined to seeds and furthermore, to specific stages of seed development (Higgins, 1986). Functional analysis of the 5' regions of genes for seed proteins from legumes and other plants has shown that these sequences confer seed-specific and temporally regulated expression within transgenic tobacco and petunia (Goldberg et al., 1989; see Table 1.1). This indicates that both the DNA regulatory elements and the transcription factors which control seed storage protein expression are conserved between different plant species.

Sequence comparisons of the proximal 5' flanking regions of 7S and 11S storage protein genes from legumes have identified some common sequences. In addition to the ubiquitous CAAT and TATA boxes, a variety of consensus sequences have been found in the 5' flanking regions of the legume seed storage protein genes including the octanucleotide (Rerie et al., 1990), CACA (Goldberg, 1986) and CATGCATG (Hoffman and Donaldson, 1985; Dickinson, 1988; Nielsen et al., 1989) elements. Longer conserved sequences; the legumin (Baumlein et al., 1986) and vicilin boxes I (Gatehouse et al., 1986) and II (Newbigin, 1988), have been found in the 11S and 7S genes, respectively. The legumin box contains a CATGCATG sequence whilst vicilin box I contains both octanucleotide and CACA sequences.

Despite the conservation of these sequences within 11S and 7S genes, promoter analysis has shown that the regions containing these sequences are not sufficient for detectable levels of gene expression. A 5' deletion series of the pea legumin gene LegA1 (Rerie, 1989; Rerie et al., 1991) showed that, although the first 237 bp of the upstream region of LegA1 contained a legumin, CAAT and TATA box, these nucleotides alone were insufficient to drive expression of the gene in transgenic tobacco. The first 668 bp of the upstream region, however, did confer detectable gene expression, indicating the presence of an upstream activating sequence located downstream of position -668 and -237.
Sequence comparisons between this portion of the LegA1 gene revealed the presence of a 23 bp AT-rich sequence, designated legumin box II (L2), that was also present in another pea 11S gene, LegA2 (Rerie et al., 1991).

Further examination of the LegA1 and LegA2 genes showed that each gene contains another conserved motif, legumin box II', located near to, and very similar to legumin box II (Figure II.1). AT-rich repeats resembling legumin box II and II' are also located in the 5' region of a pea vicilin gene, Psa-v (Figure II.1), which suggested that this sequence may be conserved, perhaps as pairs of repeats, within pea seed storage protein genes.

It was also shown that a fragment containing the conserved sequences of Psa-v did not confer high-level gene expression on the chloramphenicol acetyl transferase (CAT) gene in transgenic plants (Newbigin 1988; Chapters 2 and 4). Although the region -228/-49 of Psa-v (pEN-12) contains both of the conserved vicilin boxes and a CATGCATG sequence, this was not sufficient to confer detectable levels of gene expression in seeds. The full-length 5' region of Psa-v, -2500/-49 (pEN18), however, was capable of promoting high levels of expression within transformed tobacco seed (Newbigin, 1988), indicating, like LegA1, that additional quantitative elements must be present upstream of the conserved sequences.

It was shown that the conserved elements in the promoter of Psa-v contributes to gene expression but that enhancement by additional quantitative elements was required for high-level expression (Chapter 2). Mutational analysis indicated, however, that the CATGCATG element (sequence tATGCATG) or sequence in vicilin box I might repress expression in non-seed organs (Chapter 4). Other data (Chen et al., 1988; Bustos et al., 1991) indicate that such sequences do not inhibit expression in non-seed organs. Thus the quantitative element in seed storage genes may also contribute to seed specificity. To test this, two sequences, of similar size, were ligated to the construct pEN12, which was not expressed at detectable levels in transgenic tobacco. In the first construct, pEN21, the enhancer region from the constitutively expressed CaMV 35S promoter was used (Benfey and Chua, 1989). In the second, pEN22, a fragment containing the quantitative region from the LegA1 legumin gene (-668 to -237, Rerie et al., 1989) was inserted upstream of -228 of the vicilin 5' sequence. These gene constructs were
Conserved AT-rich DNA sequences located in upstream regions of seed genes.

The nucleotide sequences of legumin box II (53) and legumin box II' that have been found in the upstream regions of the *LegA1* and *LegA2* legumin genes are shown. A similar sequence, which was found in the 3' ends of the 55 bp repeats of the vicilin upstream region, is also shown. Gaps in the sequences (-) were included to maintain homology.
### A/T rich sequences

The quantitative elements in Psa-v, located between -775/425 and -425/627 each contain an A/T-rich repeat sequence which contained the leg box II-like sequence. Gel-shift assays were used to show that nuclear proteins from pea

<table>
<thead>
<tr>
<th>Motif</th>
<th>Sequence</th>
<th>Position</th>
</tr>
</thead>
<tbody>
<tr>
<td>Legumin box II, LegA1 gene:</td>
<td>GT - - AATT -- AA CAC TTTAA TTTGA AG</td>
<td>-350</td>
</tr>
<tr>
<td>Legumin box II, LegA2 gene:</td>
<td>GT - - AATT CAAG CAC TTTAA TTTGA AG</td>
<td>-470</td>
</tr>
<tr>
<td>Legumin box II', LegA1 gene:</td>
<td>GTA - AATT -- AG CAC TTTAA ATGTA A A</td>
<td>-580</td>
</tr>
<tr>
<td>Legumin box II', LegA2 gene:</td>
<td>GCA - AAT - - - - CAG TTTAA AAAGTA A</td>
<td>-710</td>
</tr>
<tr>
<td>55 bp repeat, Psa-v gene:</td>
<td>g^TA\textsuperscript{a}^g AATT -- A - CAT TTTA\textsuperscript{c}</td>
<td>-390, -480</td>
</tr>
</tbody>
</table>
transferred into tobacco, and assayed for expression in seeds, roots, and leaves.

The quantitative elements in *Psa-v*, located between -775/-425 and -424/-307 each contain an AT-rich repeat sequence which contained the leg box II-like sequences (Chapter 2). This suggested that the *cis*-acting element(s) that is located downstream of -668 of *LegA* may be present in the leg box II sequence. Gel-shift assays were used to show that nuclear proteins from pea seeds could be detected binding to the quantitative region. Competition assays, however, which were used to localise the DNA-protein binding sites, indicated that the nuclear protein did not bind to leg box II, but bound to other regions, located within -549/-316.
Methods

Construction of plasmids

**PQ4**
The construct pQ4 was used as a reporter gene. This gene construct contains the 5' region of *Psa-v* from -228 to -49, ligated to the TATA box region of the CaMV 35S gene (from -45 to +1), linked to the coding region of the *cam* gene (encoding the enzyme chloramphenicol acetyl transferase [CAT]) and the 3' region from the nopaline synthetase gene (in brief: *Psa-v*-228/-49-TATA-CAT-NOS). It was made by ligating the *Sma* I - *Hind* III fragment of the pUC "version" of the binary construct pEN12, construct pEN4 (containing *Psa-v*-228/-49-TATA-CAT-NOS, Newbigin, 1989) into the *Sma* I - *Hind* II site of the vector pTZ19R.

**pEN 21**
The enhancer region from the constitutively expressed CaMV 35S promoter (Benfey and Chua, 1989) was ligated as a *Hinc* II - *Eco* RV fragment (-426/-90) from p35SKN into the *Sma* I site of pUC8, generating plasmid pQ1. A fragment containing the enhancer region was then ligated as a *Sal* I - *Eco* RI (bluntended) fragment into the *Sal* I - *Sma* I sites of pQ4, generating the plasmid pEN21 (Figure II.2).

The binary construct (g)pEN21, was then constructed by ligating the *Kpn* I - *Hind* III fragment of pEN21 into the same sites in the binary vector pGA482 (An, 1987a). This gene construct, therefore, contained the 35S enhancer, ligated to the fragment containing vicilin boxes I and II, the 35S TATA box, gene encoding CAT and nos 3' end (in brief, referred to as CaMV 35S-vicilin-CAT). This construct is referred to as pEN21 in the remainder text, for consistency with the original pEN12 and pEN18 constructs (Figure II.2).
pEN22
The putative enhancer fragment from the LegA1 legumin gene was ligated as a 
SspI - NcoI (bluntended) fragment (-668/-237) into the SmaI site of pUC8, 
generating plasmid pQ3. A fragment containing the putative quantitative region 
was then ligated as a SalI - EcoRI (bluntended) fragment into the SalI -SmaI 
sites of pQ4, generating the plasmid pEN22.

The binary construct (g)pEN22 was then constructed by ligating the KpnI - 
HindIII fragment of pEN22 into the same sites in the binary vector pGA482 (An, 
1987a). This gene construct, therefore, contained the putative LegA1 enhancer 
element, with legumin boxes II and II' ligated to the fragment containing vicilin 
b oxes I and II, the 35S TATA box, gene encoding CAT and nos 3' end (in brief 
referred to as legumin-vicilin-CAT). Construct (g)pEN22 is referred to as 
pEN22 in the remainder of the text, for consistency with the original pEN12 and 
pEN18 constructs.

Genes pEN12 and pEN18 were kindly provided by Ed Newbigin; p35SKN by 
Danny Llewellyn.

Transformation of plants.

The binary plasmids containing 5' deletions of Psa-v were transferred from 
Escherichia coli (HB101) into A. tumefaciens (strain LBA4404 [Hoekema et al., 
1983]) by triparental mating, using the helper plasmid RK2013, as described by 
Ditta et al. (1980). Transconjugants containing the binary vector were selected 
on LB MG medium (20 g/l, agar; 5 g/l, Bactotryptone; 2.5 g/l, Bacto yeast 
extract; 87 mM, NaCl; 27.5 mM, mannitol; 7.82 mM, glutamic acid; KH2PO4, 
1.84 mM; 0.41 mM, pH 7.0 MgSO4-7H2O) containing rifampicin (50 mg/l) and 
tetracycline (2 mg/l).

The A. tumefaciens transconjugants were used to transform tobacco leaf tissue, 
and plants were regenerated from transformed cells, essentially as described 
in Higgins et al. (1988). In brief; leaf pieces were soaked in the Agrobacterium 
solution for 2 to 5 minutes and were then transferred to hormone free MS 
medium (revised basal medium, Murashige and Skoog, 1962): inorganic salts 
[mM]: 41.2 NH4NO3; 18.8 KNO3; 3.0 CaCl2·2H2O; 1.5 MgSO4·7H2O; 1.25 
KH2PO4; 0.2 Na2-EDTA; 0.1 FeSO4·7H2O; 0.1 H3BO3; 0.1 MnSO4·4H2O;
Figure II.2

Chimeric CAT reporter genes designed for the analysis of vicilin and legumin upstream DNA sequences.

Genes were constructed using a basic CAT reporter gene which consisted of the coding region of a chloramphenicol acetyl transferase gene and the 3' sequences from a nopaline synthase gene (Nos 3'). Portions of the promoter from the cauliflower mosaic virus (CaMV) 35S gene containing the enhancer (-420 to -91), CAAT (-90 to -46) and TATA (-45 to +1) elements were used. The TATA box sequence from the CaMV 35S gene was used for all genes, except the promoterless control, pGA492.

Upstream sequences from vicilin, legumin and the CaMV 35S were ligated in front of this basic gene. Sequences -228 to -49 of the vicilin gene contained both conserved vicilin boxes I and II (pEN12). Conserved and repeated sequences in the 5' sequence of the Psa-v gene are shown; from left to right: AT-rich repeats ( ), a palindrome ( ), vicilin box II ( ), vicilin box I (overlined, containing an octanucleotide box, , and a CACA element, ) and CATGCATG elements ( , sequences TATGCATG and ggTGCATG, respectively ) and a CAAT box (C). Positions are relative to the start site of transcription.

Legumin -668 to -237 contains A/T rich sequences legumin box II and II' (pEN22). The legumin box II and legumin box II' in LegA1 are indicated as black and stippled boxes, respectively.

The coding region for the gene coding for chloramphenicol acetyl transferase (CAT) is shown as a hatched rectangle, whilst the 3' region of the nopaline synthetase gene is shown as a black rectangle.
The next day, the leaf samples were transferred to MS medium containing indole-3-butyric acid (IBA) 1.0 mg/l, 0.005 mg/l NAA, 0.005 mg/l 2,4-D and 3% sucrose to select the transformed plants. The leaf segments were harvested from intact seedlings 4 weeks after the transformation. Each plant was grown in a 3-litre pot under greenhouse conditions for 4 weeks, and the presence of the transgenic plant was confirmed by RT-PCR. The CAT and β-glucuronidase (GUS) activities were measured in the transgenic plants. The CAT activity was assayed by measuring the conversion of 14C-p-chloromercuribenzoate to the chlorophyll derivatives. The percentage of conversion was determined by counting the radioactivity in the chlorophyll and chlorophyll derivatives. The expression of the CAT enzyme encoded by the CAT gene was measured by the conversion of 14C-p-chloromercuribenzoate to the chlorophyll derivatives.
0.03 ZnSO4·7H2O; 0.005 KI; 0.001 Na2MoO4·2H2O; 0.0001 CuSO4·5H2O; 0.0001 CoCl2·6H2O and organic substances [mg/l] 30 000 sucrose; 10 000 agar; 100 myo-Inositol; 2.0 glycine; 0.5 nicotinic acid; 0.5 pyridoxine-HCl; 0.1 thiamine.HCl).

The next day, the leaf pieces were transferred to MS9 plates containing shoot induction medium (MS medium with indole acetic acid at 0.5 mg/l, benzyl-aminopurine at 1 mg/l) and the antibiotics cefotaxime (300 µg/ml, to kill the Agrobacterium) and kanamycin (100 µg/ml, to select the transformed plant cells). Shoots formed within 2 to 3 weeks; these were transferred to MS medium containing the same antibiotics, but no hormones, for rooting. Plants that formed roots were transferred to soil and grown in the glasshouse. Seed was harvested from mature seed capsules.

Southern blots experiments were done, as described in Chapter 2.2.3, to verify the presence of the trans-genes in the putative transformants (data not shown).

Plant 35SC was kindly provided by Danny Llewellyn.

**CAT Assays**

The expression of the CAT gene in these transformed plants was analysed by measuring the activity of the CAT enzyme in various organs and at different stages of plant development. CAT assays were essentially as described by Newbiggin (1988) and used protein extracted from 10 seeds (approximately 200 µg protein).

Expression of the CAT enzyme encoded by the CAT gene was measured by the conversion of 14C chloramphenicol to acetylated chloramphenicol products; these were separated and visualised by thin layer chromatography and autoradiography. The percentage of conversion was quantified by cutting out the regions which contained the radioactive chloramphenicol and its products and measuring the amount of radioactivity present in each.
DNA fragments

DNA fragments for radioactively labelled probes and for competitor DNA were also made and isolated as described in Chapter 3.2. The legumin fragment was isolated as a *Hind* III - *Eco* RI fragment from pQ3, which contained -668 to -237 of LegA1. Sub fragments of this sequence were made by digestion at the *Msp* I (-549) and *Alu* I (-315) sites.

A smaller synthetic fragment was made containing sequences -372/-350 of LegA1, with *Pst* I and *Bam* HI sites added to the 5' and 3' ends, as shown in the inset below. To generate the legumin box II fragment, a pair of complementary oligonucleotides were made, as shown below. The oligos were synthesised on an Applied Biosystems DNA synthesiser. Equimolar concentrations of oligonucleotide were annealed by heating at 70°C for 5 min (to disrupt secondary structure), then incubated at 60°C for 5 min and then allowed to cool slowly to room temperature. The annealed product was purified and eluted from a 20% (w/v) polyacrylamide gel (see below).

```
Legumin Box II

coding 5'
GTAATAACACTTTAATTTGAAG

3'
ACGIt:ATT CATTAATTGTGAAATTAAACTTC

CTAG 3'
```

The fragment containing sequences between -840 to -310 of LegA2 was isolated from the plasmid pLegA2 (Rerie *et al.*, 1991) as an *Ssp* I - *Xba* I fragment.

A 322 bp *Pvu* II fragment from pUC118 was isolated and used as non-specific competitor.

All DNA fragments were isolated by electrophoresis in polyacrylamide gels. The concentration of the gels ranged from 3.5% - 20 % (w/v) polyacrylamide (19:1 (w/w) acrylamide: bis-acrylamide), depending upon the fragment size; and were run in 1 X TBE buffer, (89 mM Tris buffer, 89 mM boric acid, 20 mM EDTA, pH 8.3, 5% [v/v] glycerol). The fragments were extracted from gel slices using two volumes of gel elution buffer (500 mM ammonium acetate, 10 mM magnesium acetate, 1 mM EDTA, 0.1 % (w/v) SDS, pH 7.0), and incubating at 65°C for 10 min. Elution was continued at 37°C for 8 hours or 55°C for 1 hour.
and the DNA was precipitated from the eluate by the addition of two volumes of ethanol.

**Radioactive labelling of DNA fragments**

The probe for gel-shift assays was prepared by labelling the ends of the DNA fragment with DNA polymerase I (Klenow fragment), using [α-32P] dATP and [α-32P] dCTP. The labelled probe was then purified by electrophoresis in polyacrylamide gels, as described above, and diluted to 4000 cpm (Cerenkov)/µl (0.01 to 0.20 ng/µl) using sterile distilled water.

**Binding of the nuclear protein extract to the radioactive probe**

The protocol was based upon that described by Tokuhisa et al. (1990). Nuclear protein, 1.25 µg, was incubated with end-labelled fragments (4000 to 8000 cpm) for 20 min at 30°C, containing 1 µg poly (d(dC)), 5% (v/v) glycerol, 5 µg BSA, in a final volume of 20 µl binding buffer (10 mM Tris, pH 7.5, 50 mM NaCl, 1 mM EDTA, 1 mM DTT). After incubation, 1 µl of 0.1% (v/v) bromophenol blue was added to each reaction tube.

Competition assays were done as described above, except that the competitor DNA was incubated with protein extract for 10 min at room temperature prior to the addition of the DNA probes. Unless otherwise stated, the amount of competitor DNA used was a 10-fold molar excess of the probe DNA used.

**Polyacrylamide gel electrophoresis**

Reaction mixtures were electrophoresed at 4°C for 2-4 hours on 4% (w/v) polyacrylamide (40:1) gels, containing 5% (v/v) glycerol, in 0.25 x TBE buffer at 100V (approximately 20 mA), with buffer recirculation. Following electrophoresis, gels were dried and either exposed to X-ray film for 0.5 to 8 days.
Results

Functional analysis of 5' regions from vicilin and legumin genes

To investigate whether the region -228/-49 of *Psa-v* contains elements involved in seed-specific expression, and whether the quantitative region of *LegA1* -668 to -237 ([Rerie et al., 1991](#)) could activate expression of the *Psa-v* conserved sequences, two chimeric genes were constructed (Figure 11.2) and transferred to tobacco. Both of these genes contained an upstream sequence ligated to the original, non-expressing (reporter gene) construct, pEN12. The enhancer region from the CaMV 35S promoter (Benfey {59}) was added to pEN12 to make construct pEN21. The putative quantitative element from *LegA1* was added to pEN12 to make construct pEN22.

Plants transformed with a promoterless CAT reporter gene (pGA492) were the negative controls. Plants transformed by pEN18 (*Psa-v* -2500/-49-TATA-CAT) and 35SC (CaMV 35S promoter-CAT) were used as positive controls for seed-specific and constitutive expression, respectively, of the CAT gene.

Figure 11.3A shows CAT activity in the mature seed of twelve tobacco plants independently transformed with pEN21 (CaMV 35S-vicilin-CAT) plants. Five of these plants expressed CAT activity well above background levels. the remainder of the plants had CAT activity that appeared slightly higher than background activity. The variation in expression may have been due to copy number or the location of the *trans*-genes in the host chromosomes. Nevertheless, addition of a "constitutive" (CaMV 35S) enhancer region to the pEN12 gene construct resulted in expression of the CAT gene in the seed.

CAT activity was also detected in the leaves and roots of the positive pEN21 plants (Figure 11.4), indicating that the presence of the conserved vicilin sequences did not prevent gene expression in non-seed tissues.
Figure II.3

**CAT activity in the seeds of tobacco transformed with chimeric CAT genes.**

CAT activity was measured in mature seeds from 12 and 10 tobacco plants transformed with (A) pEN21 (CaMV 35S-vicilin-CAT) and (B) pEN22 (legumin-vicilin-CAT), respectively. CAT assays were essentially as described by Newbigin (1988) and used protein extracted from 10 seeds (approximately 200 µg protein). Lanes C and 35SC show the activity of CAT in seeds from plants transformed by the pGA492 (promoterless-CAT) and 35SC (CaMV 35S promoter-CAT) genes, respectively. Other lanes refer to individual plants transformed by (A) pEN21 and (B) pEN22 constructs. Chl, 1-Ac and 3-Ac refer to the substrate, $^{14}$C chloramphenicol and two acylated products, 1- and 3- acetyl chloramphenicol.
Figure 1B shows the expression of CAT activity in the mature seed of 10 independently transformed pEN21 (legumin-vicilin-CAT) plants. Three plants had detectable, but low, CAT activity. Expression from pEN22 was also low, compared to that observed in shoots from pEN18 and perfect plants (Figure 1B). However, in most plants, the relative quantitative region from L3 (60-65°C) to L4 (67°C) did not differ significantly. Expression from pEN19 (increased gene expression, although not as highly as in the perfect plant) was higher than that observed in the pEN22 plants. The data for the pEN22 plants containing the same 35SC enhancer element as in all previous experiments were included for comparison.

B
pEN22 plants

legumin-vicilin-CAT

Detection of nuclear proteins binding within a quantitative element legumin gene

In gel retardation assays, the legumin quantitative element (35SC-35SC) interacts with nuclear proteins from shoot suspensions to form two complexes.

All the solutions were prepared as described above, except that the samples were not incubated prior to electrophoresis. Two sets of experiments were performed, one at 5°C (data not shown) and one at 37°C. Each experiment was performed in triplicate, and all the results were similar. The data for the pEN22 plants containing the same 35SC enhancer element as in all previous experiments were included for comparison.

3-Ac
1-Ac
Chl

35SC C 1 2 4 7A 7B 10 11 12 15 13
Figure II.3B shows the expression of CAT activity in the mature seed of 10 independently transformed pEN22 (legumin-viciilin-CAT) plants. Three plants had detectable, but low, CAT activity. Expression from pEN22 was also low compared to that observed in seeds from pEN18 and pEN21 plants (Figure II.4). Therefore the putative quantitative region from LegA1 (-668 to -237) did indeed activate (or increase) gene expression, although not as highly as either the 35S enhancer or the vicilin upstream region -2500/-229.

CAT activity was detected only in the seeds, and not the leaves of plants transformed with constructs which had the upstream regions of either seed gene ligated in front of the conserved vicilin region (Figure II.4, see pEN18 [vicilin -2500/-229] and pEN22 [legumin -668/-237]). Constructs containing the enhancer from the 35S gene (pEN21 and 35SC) were expressed in all tissues examined; seeds, roots and leaves (Figure II.4). This indicates that the upstream regions from the two pea seed storage protein genes contribute to seed-specific expression and so may not be simple quantitative elements.

**Detection of nuclear proteins binding within a quantitative element from a legumin gene**

In gel-shift assays, the legumin quantitative element (-668 to -237) interacted with nuclear protein extracts from pea cotyledons to form two complexes. These DNA/protein complexes were designated UP (upper band) and LB (lower band) in Figure II.5, representing complexes that migrated slowly and quickly in the gel. A complex with very high mobility may also have been formed, migrating just above the free probe.

Specificity of binding to the legumin probe was determined by titration with specific (LegA1 -668 to -237) and nonspecific (pUC118 Pvu II fragment, 322 bp) competitor DNAs. Increasing amounts of specific competitor competed with the formation of both radioactively labelled complexes. Only a 20-fold molar excess was required to compete UB binding. The addition of nonspecific DNA (pUC118) also competed with the UB binding to the LegA1 quantitative element, although greater amounts (200-fold molar excess) were required. Thus UB had a greater affinity for the specific LegA1 sequence than for the non-specific sequence.
Figure 11.4

CAT activity in different organs of tobacco transformed with a range of chimeric CAT genes.

A. CAT activity in the mature seed. Lanes show CAT activity of tobacco plants transformed by pGA492 (promoterless CAT, lane C), pEN18 (full vicilin-CAT), pEN22 (legumin-vicilin-CAT, plant #7A, and pEN21 (CaMV 35S-vicilin-CAT, plant #48,) and 35SC (CaMV35S promoter-CAT) genes. The pEN21 and pEN22 plants that were used were both the second highest expressers of seed CAT activity for each construct (Figure 6).

B. Leaf CAT activity. Plants are the same as described in A.

C. Root CAT activity. Lanes C and 35SC as above. Lane pEN21 represents plant #5, which had the highest expression of CAT activity in the seeds of the pEN21 series.

Assays used approximately 400 µg protein extracted from seeds (20 mature seeds) and 40 µg of protein extracted from leaves and roots. Chl, 1-AC and 3-Ac refer to the substrate, 14C chloramphenicol and two acylated products, 1- and 3- acetyl chloramphenicol.
Figure II.5

Nuclear protein binding to the legumin 5' DNA sequence, *LegA1* -668 to -237.

Titration of binding to legumin probe by specific (legumin *LegA1* -668 to -237) and nonspecific (pUC118 *Pvu* II fragment, 322 bp) competitor DNA. In each reaction, except that shown in the first lane, 1.25 µg of nuclear protein extracted from pea cotyledons 13-15 days after flowering (200 to 300 mg per cotyledon pair) was incubated with 1 µg of poly dIdC, and 4000 cpm (approximately 2 ng) of the end-labelled legumin fragment in the presence of binding buffer. Competitor DNA fragments were added in amounts equal to 0-fold, 20-fold, 100-fold and 200-fold-fold molar excess. The samples were then loaded onto a non-denaturing, low ionic strength (.25x TBE), 4% PAGE and the radioactive bands were detected by autoradiography. The upper and lower bands, representing slow and fast motility DNA-protein complexes respectively, are marked with arrows. Free probe (fp), incubated without nuclear extract or competitor DNA was loaded in the far left lane of both figures.
In contrast, LB binding to the LegA element was not appreciably affected by a 200-fold molar excess of the competitor fragment. A 200-fold molar excess of the LegA fragment significantly reduced the radioactively labeled LB complex. The high concentration of specific competitor required to eliminate LB binding, however, suggests that this binding was probably due to non-specific DNA-protein interactions.

In order to further define the DNA binding preferences of the LegA element, -660 to +237 was digested in 100-fold excess of DNA (competitor DNA). A synthetic oligonucleotide containing the box II, between -328 to +353, was used to compete by a 10-fold molar excess of any other fragment (Figure 5). The competion was 100-fold molar excess of the competitive DNA. The oligonucleotide was found to inhibit the binding of the LegA element. However, the same competitor with competitive DNA, regions in -660 to +237, showed no change in binding of the LegA element.
In contrast, LB binding to the LegA1 element was not appreciably affected by a 200-fold molar excess of the nonspecific fragment. A 200-fold molar excess of the LegA1 fragment significantly reduced the radioactively labelled LegA1-LB complex. The high concentration of specific competitor required to eliminate LB binding, however, suggests that this binding was probably due to non-specific DNA-protein interactions.

In order to further define the DNA binding site(s), the legumin quantitative element, -668 to -237, was digested into three fragments. These were: -660/-550, containing leg box II; -549/-316, containing leg box II; and -315/-237. A small synthetic oligonucleotide, containing the A/T rich legumin box II, between -372/-350, was also used as competitor DNA. UB was competed by a 10-fold molar excess of -549/-316, but not by a similar excess of any other fragment (Figure II.6). Similarly, LB binding was competed by an 100-fold molar excess of the -549/-316 fragment (data not shown). This indicates that the binding sites for both BP1 and BP2 are located within -549/-316. Because the legumin box II oligo failed to compete with binding, however, the site with which UB (and perhaps LB) interacts is located at other regions in -549/-373; that is, between -549/-316 and/or -349/-316. A LegA2 fragment also failed to compete effectively with UB binding.
Figure II.6

UB does not bind to legumin box II in the quantitative element of LegA1.

Top panel: Diagram of the DNA fragments that were used as competitor DNA (a) LegA1 -668/-237, (b) LegA1 -549/-316, (c) legumin box II [LegA1 -372/-350, (d) LegA1 -668/-550 plus -315/-237 and (e) LegA2 -840/-310. Fragment (a) LegA1 -668/-237 was also used as the probe. The location of legumin box II and legumin box II' in LegA1 and LegA2 are indicated as black and stippled boxes, respectively. The region containing the UB binding site is shown inside a hatched oval.

Bottom panel: Competition of binding of AT-rich fragments from the legumin genes LegA1 and LegA2. The LegA1 probe was incubated in the absence of nuclear extract (first lane), or with 1.25 µg of nuclear extract from 13-15 DAF seeds. A ten-fold molar excess of each unlabelled competitor DNA fragment (a, b, c, d, e) was added, as indicated. No competitor DNA was added to the reactions in the second lane. All incubations contained of 1 µg of poly (dIdC). DNA protein complexes UB, LB and the free probe (fp) are indicated.
Competitor DNA

- **a** -668 to -237
- **b** -549 to -316
- **c** Legumln box II
- **d** -668 to -550, -315 to -237
- **e** Leg A2: -840 to -310

**DNA probe**

<table>
<thead>
<tr>
<th>Competitor DNA</th>
<th>LegA1 -668 to -237</th>
</tr>
</thead>
<tbody>
<tr>
<td>- - a b c d e</td>
<td>- + + + + + + +</td>
</tr>
</tbody>
</table>

**Extract**

UB

LB

fp
Discussion

Upstream (quantitative) elements contribute to seed-specific expression.

Analysis of the 5' region -228/-49 of \textit{Psa-v} showed that the sequences within this vicilin promoter, although it contains the conserved vicilin boxes I and II, were insufficient to confer detectable CAT gene expression within tobacco seeds (Newbigin, 1988), or leaves (Figure II.4). Addition of a the 35S enhancer to these conserved vicilin sequences conferred expression within transgenic seeds, but did not prevent gene expression in other organs of the plants. This indicates that these sequences did not contain an inhibitory element capable of repressing transcriptional activation in non-seed tissues.

This result differs to those for similar gene constructs with nodule-specific and other genes (Ellis \textit{et al.}, 1987; Stougaard \textit{et al.}, 1987; Stockhaus \textit{et al.}, 1989; Keil \textit{et al.}, 1990; Stougaard \textit{et al.}, 1990), where addition of the 35S enhancer fragment increased expression in the appropriate tissues, without activating expression in other tissues. The lack of tissue specificity, however, does resemble the expression of other seed-specific promoter elements when coupled with the 35S enhancer (Chen \textit{et al.}, 1988; Bustos \textit{et al.}, 1989; Fujiwara and Beachy, 1992), indicating that seed-specific gene expression may be regulated in a different fashion to other tissue-specific genes.

Expression in the seeds, however, may have been enhanced by the presence of the conserved sequences from the 5' region of \textit{Psa-v} when placed downstream of the 35S enhancer. One plant in particular, pEN21-5, had 10-20 times more CAT activity than that found in the seeds of the highest expressing 35SC plant (Figure II.4). Statistical analysis comparing the expression levels in many plants transformed with 35SGN, as well as pEN22, is required to confirm the enhancement effect, if any. The observation supports results from Chen \textit{et al.} (1988), who showed that the region -257/-78 of a soybean vicilin gene, \textit{Gma-\alpha'}, conferred 40-fold enhancement of expression within transformed seed. Only sequences between -120/-77 of the seed-specific enhancer element from \textit{Gma-\alpha'} are similar to sequences within region -228/-49 of \textit{Psa-v} (-130/-88). These similar regions contain vicilin box I, including both CACA and octanucleotide boxes. Both regions also contain CATGCATG elements,
although not in similar positions. The presence of these sequences in both constructs suggests that one or more of these DNA motifs may enhance gene expression in seeds.

Figure 11.4 show that the quantitative region from legumin conferred weak, but seed-specific expression, when ligated in front of the truncated vicilin promoter present in construct pEN12 (nucleotides -228/-49). It would be interesting to determine whether this fragment can function to activate expression in the absence of seed conserved sequences such as the vicilin boxes or legumin box, or whether it is only able to enhance gene expression by interaction with other elements.

Only gene constructs which contained upstream regions from seed-specific genes; either the legumin quantitative region or -2500/-229 of the vicilin 5' sequence, conferred seed-specific expression in transgenic tobacco. Thus, sequences involved with seed-specific expression may be present in the upstream regions as well as the promoters of seed storage protein genes. These may function either by containing elements that generally (constitutively) enhance gene expression plus elements that suppress gene expression in non-seed tissues. Alternatively, these AT-rich quantitative elements may contain cis-acting elements that are only bound and activated by transcription factors that are only active in seeds.

**Legumin quantitative element contains a DNA-protein binding site.**

Gel-shift analysis showed that nuclear proteins bind specifically within -549/-316 of LegA1. This correlates with the binding site within a homologous allele from pea, LegA (Meakin and Gatehouse, 1991). The same proteins, termed LABF-1, also bind between -833/-582 of LegA (Meakin and Gatehouse, 1991). Comparison of these regions located two similar sequences, legumin box II (Rerie, 1989) and legumin box II (Figure 11.1) within -549/-316 and -833/-582, respectively, and so it was proposed that legumin box II and II' may contain the UB/LABF-1 binding sites. The fragments containing these sequences, however, -668/-550 (legumin box II") and -372/-350 (legumin box II), failed to compete. Thus although UB1 (and perhaps LB2) binds between -549/-316, presumably between -549/-373 and/or -349/-316, it does not bind to either legumin box II or legumin box II'.
The region of *LegA2* also contains both legumin II boxes (Figure II.1). A fragment continuing both these sequences was also unable to compete (as effectively as *LegA1*-549/-316) with UB binding, confirming that the legumin box II sequences do not contain the UB binding sites. Since the UB binding was not competed by this large, AT-rich DNA fragment this also indicates that UB binds to specific sequences in the *LegA1* gene.

Comparison of the binding regions, -833/-669 and -549/-316, for similar motifs that might contain the binding sites located six motifs or 9 bp in length that were present in both sequences (not shown). Nevertheless, the binding sites may not necessarily be in very similar sequences. Nuclear proteins that bind to AT-rich sites within other genes that are highly expressed in seeds have been shown to have loose sequence requirements, indicating that these proteins may recognise a range of AT-rich sequences (Bustos *et al.*, 1989; Riggs *et al.*, 1989; Pedersen *et al.*, 1991). There is, however, obviously some specificity in binding, since not all AT-rich sequences compete as well as other AT-rich sequences (Bustos *et al.*, 1989; Riggs *et al.*, 1989; Jofuku *et al.*, 1990; Pedersen *et al.*, 1991). The AT-rich sequences -668/-550, -372/-350 and -315/-237 of *LegA1* also failed to compete with binding to -549/-316.

Short runs of A or T nucleotides (such as AAAA or TTTT) can confer an intrinsic bend on a DNA molecule, which suggests that such bends, when part of a protein binding site, might facilitate the binding of that protein to the binding site (Travers, 1989). Thus, DNA-binding proteins that bind to the AT-rich regions in seed-specific genes may recognise and bind to particular bends and twists in the DNA molecule that could be induced by a variety of AT-rich sequences.
Further experiments are required to determine the location of the binding sites in the quantitative element of \textit{LegA1}, and also, whether these binding sites contain the \textit{cis}-acting elements that activate gene expression.
BIBLIOGRAPHY


Bruce, W.B. and Quail, P.H. (1990) cis-acting elements involved in photoregulation of an oat phytochrome promoter in rice. Plant Cell 2, 1081-1089


Datta, N. and Cashmore, A.R. (1989) Binding of a pea nuclear protein to promoters of certain photoregulated genes is modulated by phosphorylation. Plant Cell 1, 1069-1077


Dehesh, K., Bruce, W.B. and Quail, P.H. (1990) A trans-acting factor that binds to a GT-motif in a phytochrome gene promoter. Science 250, 1397-1399


Descombes, P. and Schibler, U. (1991) A liver-enriched transcriptional activator protein, LAP, and a transcriptional inhibitory protein, LIP, are translated from the same mRNA. Cell 67, 569-579


Donald, R.G.K. and Cashmore, A.R. (1990) Mutation of either G box or I box sequences profoundly affects expression from the Arabidopsis rbcS-1A promoter. EMBO J. 9, 1717-1726


Holdsworth, M.J. and Laties, G.G. (1989b) Site-specific binding of a nuclear factor to the carrot extensin gene is influenced by both ethylene and wounding. Planta 179, 17-23


Jofuku, K.D. and Goldberg, R.B. (1989) Kunitz trypsin inhibitor genes are differentially expressed during the soybean life cycle and in transformed tobacco plants. Plant Cell 1, 1079-1093


Osborne, T.B. (1924) The vegetable proteins, Longmans, Green, London


Schmidt, R.J., Ketudat, M., Aukerman, M.J. and Hoschek, G. (1992) Opaque-2 is a transcriptional activator that recognizes a specific target site in 22-kD zein genes. Plant Cell 4, 689-700


Shirsat, A.H., Meakin, P.J. and Gatehouse, J.A. (1990) Sequences 5' to the conserved 28 bp Leg box element regulate the expression of pea seed storage protein gene legA. Plant Mol. Biol. 15, 685-693


Spiker, S. (1988) Histone variants and high mobility group non-histone chromosomal proteins of higher plants: Their potential for forming a chromatin structure that is either poised for transcription or transcriptionally inert. Physiol. Plant. 75, 200-213


Tabata, T., Nakayama, T., Mikami, K. and Iwabuchi, M. (1991) HBP-1a and HBP-1b: leucine zipper-type transcription factors of wheat. EMBO J. 10, 1459-1467


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

Voelker, T., Sturm, A. and 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


Wandelt, C.I., Khan, M.R.I., Craig, S., Schroeder, H.E., Spencer, D. and Higgins, T.J.V. (1992) Vicilin with carboxy-terminal KDEL is retained in the endoplasmic reticulum and accumulated to high levels in the leaves of transgenic plants. Plant J. 2, 181-192


Summary Figure

Conserved DNA motifs, regulatory elements and DNA-protein binding sites in the 5' region of the Psa-v gene.

Conserved and Repeated Sequences
The location of conserved and repeated DNA sequences in the region between -773 to +1 of the 5' sequence of the Psa-v gene are shown; from left to right: AT-rich repeats (~), vicilin box II (▼), vicilin box I (overlined, containing an octanucleotide box, ▼), and a CACA element, ▼), CATGCATG elements (△ and Δ, sequences tATGCATG and ggTGCATG, respectively), a CAAT box (C), and a TATA box (T). Positions are relative to the start site of transcription.

Regulatory elements
The location of regulatory elements between -773 and +1 are shown, as defined in Chapters 2 and 4. The number of crosses indicates the relative level of expression conferred by these elements. The minus indicates a negative element. "Seed" indicates a region that inhibited expression in leaves and petals and increased expression in seeds.

Binding of Nuclear Proteins
Location of DNA-protein binding sites between -773 and +1 of the Psa-v gene, Chapter 3. AT-2 binding sites are shown as green triangles. Regions required for TEMP-I are shown as red semicircles, with the (putative) consensus binding sequences in forward (●) and reverse (○) orientation. TEMP-2 and TEMP-3 sites, both located in -61/-4 are represented by yellow and empty arches (∩, ∪).