Cloning and characterisation of cDNAs encoding the major, pea storage proteins, and expression of vicilin in E.coli

Ashton Joseph Delauney

TO MY PARENTS.

CLONING AND CHARACTERISATION OF cDNAs ENCODING THE MAJOR, PEA STORAGE PROTEINS, AND EXPRESSION OF VICILIN IN *E.coli*.

A Thesis Submitted by :

ASHTON JOSEPH DELAUNEY

In accordance with the requirements for the degree of Doctor of Philosophy in the University of Durham.

Department of Botany.

December, 1984

Cloning and characterisation of cDNAs encoding the major, pea storage proteins, and expression of vicilin in *E.coli*

ASHTON JOSEPH DELAUNEY

ABSTRACT

A cDNA library was constructed using mRNA from developing seeds of pea (*Pisum sativum* L.). Clones encoding legumin and vicilin, the major storage proteins, were isolated and characterised, in several cases to the extent of complete DNA sequencing.

The composite DNA sequence of the two longest legumin cDNAs extended over almost 90% of a complete legumin gene coding sequence. Both these clones contained three ∿54bp tandem repeats in the region encoding the acidic subunit. Evidence is presented that these repeats may be present in all chromosomal legumin genes and consequently, that the absence of the repeats from a previously isolated legumin cDNA probably represented a cloning artefact.

Two, near full-length, vicilin cDNAs encoding 50000-*M*r vicilin subunits, and another encoding a 47000-*M*r subunit were sequenced. The 50000-*M*r vicilin cDNAs were almost identical over most of their lengths, but one contained an artefactual, inverse repeat at its 5' terminus, while sequence differences at the 3' termini indicated the use of alternative polyadenylation sites. Comparisons of protein and cDNA-encoded amino acid sequences indicated that vicilins are synthesised as polypeptide precursors which subsequently undergo the removal of an N-terminal signal peptide and possibly a C-terminal extension, as well as being susceptible to endo-proteolytic processing. Extending these comparisons to legumin and lectin sequences suggests that endo-proteolysis of these seed proteins occurs on the C-terminal side of asparagine residues located within β-turn conformations in hydrophilic regions of the proteins.

Two vicilin cDNAs were expressed as both fused and unfused products in *Escherichia coli* under the influence of the phage lambda, leftward promoter (λP_L). Levels of expression obtained with different expression plasmid constructions supported previous hypotheses that translational efficiencies were lowered when the Shine-Dalgarno sequence was sequestered into double-stranded regions of the mRNA. There was also some indication that synthesis of a vicilin polypeptide bearing a signal peptide had a deleterious effect on the viability of the host strain.

CONTENTS.

Contents Contd/....

Page:

FIGURES.

Figures Contd/..... Page:

TABLES

ACKNOWLEDGEMENTS.

I wish to thank my supervisor, Professor Donald Boulter, for his frequent discussions, advice and encouragement throughout this study.

I am especially grateful to Dr. Ronald Croy who has been an invaluable source of help and inspiration during the past three years. The present work owes much to his careful guidance.

Special thanks are due to Mrs. Phillipa Brown who introduced me to a range of techniques at the beginning of my study; to Dr. Grantley Lycett who taught me DNA sequencing; and to Mr. Paul Preston who demonstrated the techniques of SDS-polyacrylamide gel electrophoresis and Western blotting. Mrs. Margaret Richards and Mr. Tony Pickard also contributed excellent and much appreciated technical assistance. Other members of this department, particularly Dr. Marta Evans, Dr. Anil Shirsat, Dr. John Gatehouse, Mr.David Bown and Miss Roslyn Sawyer, deserve thanks for their interest, useful discussions, and practical help whenever possible.

Several people helped in the preparation of this thesis. My thanks go to Dr. Croy and Professor Boulter for critically reading the manuscript; to Mr. Tom Hall for preparing, at short notice, the figures illustrating the construction of expression plasmids; to Mr. Paul Sydney for photography; and to Mrs. Joan Scott for typing the manuscript.

This work was supported by funds from Genzyme Biochemicals Ltd., and an Overseas Research Students Award from the Committee of Vice-Chancellors and Principals.

MEMORANDUM.

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

Lycett, G.W., Delauney, A.J. and Croy, R.R.D. (1983). Are plant genes different? FEBS Lettr. 153, 43-46.

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

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

Lycett, G.W., Delauney, A.J., Zhao, W., Gatehouse, J.A., Croy, R.R.D and Boutler, D. (1984). Two cDNA clones coding for the legumin protein of *Pisum sativum* L. contain sequence repeats. Plant Mol. Biol. 3, 91-96.

Lycett, G.W., Croy, R.R.D., Delauney, A.J., Shirsat, A., and Boulter, D. (1984) Molecular analysis of the gene families coding for the storage proteins of *Pisum sativum* L. Heredity, in press.

ABBREVIATIONS.

Abbreviations are used as recommended in the "Biochemical Journal Instructions to Authors" (Biochemical Society, 1975), with the additions listed below.

The one-letter notation for amino acids is given in "The Biochemical Journal (1969), Vol.113, pp 1-4."

bp : base pairs

Kb : Kilobase pairs

cDNA : Complementary DNA

ds-DNA : Double stranded DNA

ss-DNA : Single stranded DNA

dd-NTP : Dideoxynucleoside triphosphate

c.p.m. : Counts per minute.

EtdBr : ethidium bromide.

poly$(A)^+$ RNA : polyadenylated RNA

SDS : Sodium dodecyl sulphate.

PAGE : Polyacrylamide gel electrophoresis.

BSA : Bovine serum albumin.

SSC : Saline sodium citrate (0.15M NaCl, 0.015M Sodium citrate pH7.0)

$\lambda O_L P_L$: leftward operator-promoter region of phage lambda

N-terminal : amino terminus of a protein.

C-terminal : Carboxy terminus of a protein.

5' : 5' terminal phosphate in a DNA or RNA molecule.

3' : 3' terminal hydroxyl in a DNA or RNA molecule.

1. INTRODUCTION.

1.1. General Introduction.

Agriculture is the exploitation of the ability of crop plants to convert simple nutrients into products that are readily assimilable by man and his livestock. Cereals and legumes are the most important food crops, their seeds providing about 70% of the dietary protein of humans (Oram and Brock, 1972). Man's dependence on these crops is even greater, considering that farm animals are extensively fed on seed meals. Yet, despite their undisputed nutritive value, cereal grains and legume seeds are not ideal sources of proteins for monogastric animals since they are, with few exceptions, deficient in certain essential amino acids. The storage proteins of legume seeds are generally deficient in sulphur amino acids, whereas cereal proteins are deficient in lysine, threonine and tryptophan (Shewry *et al.*, 1981.) An improvement in the nutritional value of seed storage proteins to make them better suited to the dietary requirements of humans and other monogasts such as pigs and poultry, is therefore highly desirable. Additionally, there is an urgent need for an increase in overall agricultural productivity to meet the requirements of the ever growing world population. The latter objective may be acheived either by improving the methods of exploiting the photosynthetic capacity of crop plants, or by improving their innate performance.

The first of these measures has been successfully adopted over the past few decades in the highly industrialised countries where modern, intensive farming practices have resulted in steady increases in crop yields. However, such capital-intensive technology, with its dependence on artificial fertilisers, growth regulators, herbicides, pesticides, massive inputs of energy and a sophisticated industrial base, is unlikely to be widely applicable in economically underdeveloped countries where the shortfall between food supply and demand is most pronounced. To feed the expanding population of the Third World, it will be necessary to improve, both qualitatively and quantitatively, the intrinsic productivity of crop plants.

In fact, since man first harvested cereals for food some 17-18300 years ago (Wendorf *et al.*,1979), considerable success has been achieved in breeding new improved varieties of crop plants, the most spectacular being that which brought about the so-called "green revolution" of recent years. Conventional plant breeding involves the crossing of several variants followed by screening of the progeny for improved phenotypes. The technique is an empirical exercise based on principles established largely by trial-and-error. Though it has been proved to be extremely powerful and will, no doubt, continue to be extensively used, it is hampered by the fact that the selection of improved progeny occurs at the phenotypic level, with neither a precise understanding nor control of the underlying molecular mechanisms. Consequently, when a desirable trait has been successfully bred, it is often accompanied by other undesirable characteristics. Traditional plant breeding programs also suffer from the basic biological constraint that only sexually compatible cultivars can be crossed. Moreover, extensive inbreeding narrows the genetic base of widely cultivated crop plants with the concommitant risk that genes for important traits such as pathogen resistance and stress tolerance might be permanently lost.

The advent of recombinant DNA technology promises to revolutionize plant breeding programs. It is anticipated that, using genetic engineering techniques, it will eventually be possible to transfer to crop plants the specific gene, or genes, responsible for desirable phenotypes. The transferred genes might originate from sexually incompatible species and even, conceivably, from animals and micro-organisms. Thus, genetic diversity stands to increase, rather than decrease, as a result of the application of genetic engineering methodology to the improvement of crop plants.

Several strategies for improving crop productivity have been widely discussed as being amenable to the genetic engineering approach. These include increasing resistance to plant pathogens and herbicides, widening the range of plant species with the ability to fix nitrogen, increasing stress tolerance, enhancing photosynthetic capacities and improving the nutritional quality of plant proteins. The potential for accomplishing these objectives has been reviewed recently (Croy and Gatehouse, 1985; Barton and Brill, 1983; Larkins, 1983; Shewry *et al.*,1981); what emerges is that the prospects for eventual success are hindered by two important factors.

The first is a technical barrier. It has been possible for some time now, to introduce foreign DNA sequences into plant cells via transformation vectors derived from *Agrobacterium* Ti-plasmids (e.g. Matzke and Chilton, 1981; Leemans *et al.,* 1981; Barton *et al.,* 1983; Shaw *et al.,* 1983) but it is only in the past eighteen months or so that phenotypic expression of foreign genes in transformed plants has been achieved (reviewed by Shaw,1984). First Ti-plasmid vectors containing chimaeric, bacterial antibiotic resistance genes which functioned as dominant, selectable markers in transformed plant cells were constructed (Herrera-Estrella *et al.,*1983 a,b; Fraley *et al.,* 1983; Bevan *et al.,*1983). Using similar vectors, expression of a bean phaseolin gene in transformed sunflower cells (Murai *et al.,* 1983) and light-regulated expression of a pea ribulose-1,5-bisphosphate carboxylase small subunit gene in transformed petunia cells (Broglie *et al.,*1984) have subsequently been obtained. However, a number of problems concerning plant transformation remain unresolved. One limitation is the narrow host range specificity of *Agrobacterium* species. Ti-plasmid vectors can, so far, only be used to transfer genes to dicotyledonous plants, since monocotyledons, including many major crop plants, are not susceptible to *Agrobacterium* infection (Flavell and Mathias, 1984). Another problem is that initial transformation experiments are performed at the level of the single protoplast, and if genetic engineering techniques are to have any great impact on plant breeding, it will be necessary to regenerate healthy whole plants, from cultured transformed cells. Horsch *et al.* (1984) and De Block *et al.* (1984) were able to generate morphologically normal and fertile plants from tobacco cells transformed with Ti-plasmid-derived, antibiotic resistance vectors, but unfortunately, many agronomically important plants particularly the monocot crop species have proved refractory to regeneration from cell culture (Flavell and Mathias, 1984). Nevertheless, none of these problems appear to be insurmountable. The development of hybrid Ti-plasmid : gemini virus vectors has been proposed as a means of circumventing the limited host specificity of *Agrobacterium* (Buck and Coutts, 1983). The use of vectors incorporating plant transposable elements (Fedoroff, 1983) in a manner analagous to the transformation of *Drosophila* with P-element-derived vectors (Rubin and Spradling, 1982) is another promising avenue to be explored. Vectors based on the Ri-plasmid of *Agrobacterium rhizogenes* might enable easier regeneration of whole plants from transformed cells (Chilton *et al.,*1982).

There is clearly no shortage of ideas, and it seems reasonable to expect that this very active area of research will yield refinements of current procedures and solutions to presently unresolved difficulties.

A more formidable barrier to the implementation of genetic engineering techniques into plant breeding is the fact that little is understood at present about the physiological and molecular processes underlying many of the traits that the plant breeder might wish to alter. Given the recent advances in plant transformation technology, it is likely that molecular biologists will be able to routinely obtain expression of foreign genes in many of the important crop plants, before it is known what genes might be gainfully transferred to these plants.

One exception to this generalization is the prospect of being able to improve the nutritional quality of food crops by manipulating the storage protein genes of these crops. Considerable effort has been devoted, since the beginning of this century, to the characterisation of storage proteins and to the study of their synthesis and deposition in the developing seeds of legume and cereal crops. The techniques of recombinant DNA technology have recently been added to the armoury of investigators studying these systems. As a result, an impressive body of information on diverse aspects of storage protein biochemistry has been accumulated (for recent reviews see Derbyshire *et al.*,1976; Boulter, 1981; Larkins, 1981; Brown *et al.*, 1982; Gatehouse *et al.*, 1984; Higgins, 1984; Croy and Gatehouse, 1985; and Vol.304B of Phil.Trans.R. Soc. Lond.,1984, pp. 273-407).

The impetus for the study of seed storage proteins derives largely from the nutritional and economic value of these proteins. Recently however, interest has been stimulated for a different reason. Storage protein synthesis is a strictly controlled process whereby certain tissues produce a few specific proteins in vast quantities during precise periods in the differentiation of the seed. The specificity of storage protein synthesis, being the result of temporally and spatially regulated gene expression, therefore provides an ideal system for investigating the mechanisms of gene regulation. A detailed understanding of the regulation of gene expression in eukaryotes is probably the most challenging goal of modern biology and studies have previously been concentrated on analagous systems such as haemoglobin synthesis in erythroid cells and ovalbumin synthesis in chicken oviducts (O'Malley *et al.*, 1977). Now, an increasing amount of work focusses on the developing seed as a model system. The first fruits to be borne from this area of research in plant molecular biology have been the isolation and characterization of genes encoding several different types of legume and cereal seed storage proteins (see Sorenson, 1984). Although the elucidation of the molecular mechanisms of gene regulation is still a long way off, the availability of cloned storage protein genes combined with the accumulated information on storage protein biochemistry provides the potential for improving the nutritional quality of seed storage proteins by means of genetic manipulation techniques. Before examining this potential further, it is necessary to consider the structure and properties of seed storage proteins and their genes. The following discussion will be limited to legume, particularly pea, seed storage proteins and their coding sequences.

1.2. Structure of Pea Storage Proteins.

Seed storage proteins are generally defined as proteins which (i) are synthesised only in the seed, during seed development, and usually accumulate to levels which constitute a large proportion of the total seed protein; (ii) are deposited in membrane-bound organelles—the protein bodies; and (iii) are hydrolysed during germination to provide nutrients (nitrogen, sulphur and some carbon) for the developing seedling.

Using a classification system based on the solubility of proteins in different solvents, Osborne (1924) found that legume seeds contained primarily a group of proteins extractable with 5% saline and which he categorised as "globulins". Danielsson (1949), using density gradient centrifugation to analyse the globulin fraction, showed that it consisted of two major types of protein fractions with sedimentation coefficients of ∿11-13S and 7-8S. The relative proportions of these types of proteins vary considerably within the Leguminosae : at one end of the scale, *Phaseolus vulgaris* contains very little, if any of the ∿11S protein, whereas at the other extreme, the 11S protein is the predominant storage protein in *Vicia faba* (Gatehouse *et al.*, 1984). *Pisum sativum* is a typical legume, containing approximately equal amounts of both types of proteins which together account for ∿70% of the seed protein though there is some variation among different genetic lines (Schroeder, 1982). The 11S and 7S globulins are called legumin and vicilin respectively. Both legumin and vicilin are composed of subunits which exhibit a significant amount of size and charge heterogeneity.

1.2.1. The Legumin Fraction.

The accepted structural model for legumin was originally derived from studies on the 11S protein from *Vicia faba* (Wright and Boulter, 1974), and the structure of *Pisum sativum* legumin was subsequently shown to be essentially consistent with it (Croy *et al.*,1979). The basic model proposes that legumin is a hexameric molecule of M_r 360000-400000 consisting of six subunit pairs, each of which comprises a ∿40000-M_r subunit linked by disulphide bonds to a ∿20000-M_r subunit. The larger or α-subunits have pI valves of 4.8-6.2 and are thus referred to as "acidic", whereas the smaller or β-subunits have pI values of 6.2-8.0 and are referred to as "basic" subunits (Matta *et al.*, 1981). SDS-PAGE of purified legumin under reducing conditions shows prominent bands corresponding to the 40000-M_r and 20000-M_r subunits. However, the subunit pair is regarded as the fundamental unit of the legumin holoprotein since it has been demonstrated that legumin is synthesised as a ∿60000-M_r precursor polypeptide which is subsequently cleaved to produce the smaller subunits (Croy *et al.*,1980a). The model just described is accepted as a fair approximation of the general structure of legumin-type globulins throughout the Leguminosae, but in fact, the actual structure of pea legumin is rather more complex. Matta *et al.* (1981) showed that superimposed on this simple model is a considerable degree of heterogeneity with respect to the existence of different molecular forms, the molecular weights of subunit pairs, and the molecular weights and pI values of the constituent subunits. It was suggested that the observed polypeptide heterogeneity probably resulted both from genetic heterogeneity as well as from post-translational protein modifications.

1.2.2. The Vicilin Fraction.

The pea vicilin fraction comprises two major heterogeneous protein types of *Mr* ∿ 170000 and ∿ 280000. The subunit composition of this fraction is very complex, major polypeptides of approximate *Mr*'s 71000, 50000, 33000, 19000, 16000, 13500 and 12500 being revealed by SDS-PAGE. For some time, there was confusion as to the relationship between these subunits and the holoproteins; recently, a clearer picture has emerged. The 280000-*Mr* protein, named convicilin, has been shown to be separable from the 170000-*Mr* protein by non-dissociating techniques (Croy *et al.*, 1980b; Casey and Sanger, 1980). It is thought to consist of three or four 71000-*Mr* subunits which are not disulphide-bonded although sulphur amino-acids are present in the subunits. Different convicilin subunits have different pI values but microheterogeneity appears to be less than that of legumin or vicilin(Croy *et al.*, 1980b). Vicilin itself, *Mr* ∿170000, is thought to be synthesised and assembled from three, non-disulphide-linked subunits of *Mr* ∿ 50000 (Gatehouse *et al.*, 1981). These subunits show considerable sequence heterogeneity : some of them contain up to two specific sites for proteolytic cleavage, and the smaller vicilin subunits observed on denaturing polyacrylamide gels are derived by post-translational proteolysis of susceptible 50000-*Mr* subunits (Gatehouse *et al.*, 1982; Spencer *et al.*, 1983). Pea vicilin contains small amounts of covalently linked carbohydrate, unlike legumin and convicilin which are not glycosylated (Gatehouse *et al.*, 1984, and refs. therein). Glycosylation is confined to two vicilin subunits of *Mr* 50000 and 16000; the latter appears to be a glycosylated variant of the 12500-*Mr* subunit (Gatehouse *et al.*,1984, and refs. therein). Though vicilin and convicilin are distinct proteins, they have been shown to be antigenically related since antibodies raised against vicilin react with convicilin (Croy *et al.*, 1980b). However, the degree of relatedness at the structural or sequence level was not investigated further.

1.3. Pea Storage Protein Genes.

Several research groups have applied the techniques of recombinant DNA methodology to the study of the storage protein genes of legumes and cereals (reviewed by Sorenson, 1984). Since a significant proportion of the results to be presented in this thesis consists of the cloning and characterisation of pea storage protein cDNAs, the following information will be restricted to data which were available prior to the commencement of this work and to more recent data which do not pre-empt the contents of the "Results" and "Discussion" sections.

1.3.1. Legumin genes.

Two legumin cDNAs were sequenced by Croy *et al.* (1982), the longer of which comprised ∿38% of the legumin mRNA. That cDNA contained the entire coding sequence of the 20000-*Mr* legumin β-subunit at the 3' end of the clone, and some 30 amino acid residues of the C-terminal region of the 40000-*Mr* α-subunit. There were no in-phase initiation or termination codons in the region immediately upstream of the β-subunit coding sequence which confirmed the *in vitro* translation data (Croy *et al.*, 1980a) showing that legumin subunit pairs were synthesised as 60000-*Mr* precursors. In the absence of C-terminal amino acid sequence data for the acidic subunit, it was not possible to locate the precise site of cleavage between the two subunits, though Croy *et al.* (1982) speculated that cleavage might occur at a pair of adjacent arginine residues five residues upstream of the N-terminus of the β-subunit, by analogy with the processing of certain animal hormone precursors.

Recently, four different legumin genes were isolated from pea genomic banks (Croy *et al.*, 1985). One of these genes has been completely sequenced, revealing a number of important features (Lycett *et al.*, 1984a). The gene encoded a legumin precursor which contained a 21 amino acid-long signal peptide followed by a 36440-*Mr* α-subunit and a 20190-*Mr* β-subunit. The product of this particular gene was relatively rich in sulphur amino acids, containing 3 met and 5 cys residues in contrast to the cDNA sequenced by Croy *et al.* (1982) which encoded a single methionine and a single cysteine residue out of a total of 216 residues. Two small introns, each 88bp long, interrupted the sequence encoding the α-subunit while a third intron, 99bp long, was present in the sequence encoding the β-subunit. The boundary sequences of these introns were typical of higher plant genes. The 5' untranscribed flanking region of the gene contained all the putative transcription control sequences including a "TATA" box, a "CAAT" box and an "AGGA" box, while the 3' flanking region contained three putative polyadenylation signals.

Using a cloned cDNA to probe restriction digests of pea genomic DNA, Croy *et al.* (1982) estimated that there were ∿4 copies of legumin genes per haploid genome. That figure is probably an underestimate since the hybridisation experiments were done under high stringency conditions, and results to be described later show that DNA fragments sharing 95% homology may fail to cross-hybridise significantly under these conditions. Indeed, Shirsat (1984), using a genomic legumin clone to probe pea genomic digests, calculated that there were at least 7 legumin genes in the haploid genome. Thus, the legumin proteins are encoded by a small, multigene family which probably accounts for some of the heterogeneity seen amongst the protein subunits.

1.3.2. Vicilin Coding Sequences.

cDNA clones coding for vicilin subunit precursors were also produced by Croy *et al.* (1982). Several of these clones hybrid-selected mRNA species encoding 50000-*M*r precursors, while one selected an mRNA which encoded a 47000-*M*r precursor. As noted in section 1.2.2., vicilin subunits of *M*r <50000 are derived by post-translational proteolysis of ∿50000-*M*r precursors (including the 47000-*M*r subunit). Gatehouse *et al.*(1982) were able to establish the ordering of the small vicilin subunits relative to a 50000-*M*r precursor polypeptide by comparing the protein sequences predicted from a partially sequenced 50000-*M*r cDNA with amino acid sequences determined from purified vicilin subunits.

1.3.3. Convicilin Coding Sequences.

A cDNA encoding part of a convicilin 71000-*M*r subunit has recently been cloned (Domoney and Casey,1983). Its sequence was found to share substantial homology with the sequences of vicilin cDNAs (Casey *et al.*,1984), consistent with the serological relatedness of vicilin and convicilin. However, the coding sequences were sufficiently divergent to prevent cross-selection of mRNA transcripts by cDNAs for either of the two proteins in hybrid-release translation experiments.

1.4. Genetic Engineering of Legume Storage Protein Genes.

Two recent reviews have dealt in detail with the prospects for improving the nutritional value of seed storage proteins by genetic engineering techniques (Croy and Gatehouse, 1985; Larkins, 1983). Exhaustive coverage of the subject is therefore inappropriate here and only a brief discussion of the possibilities will be presented.

As mentioned previously, the major nutritional limitation of legume storage proteins is the deficiency of methionine and cysteine. Using techniques for *in vitro* site-directed mutagenesis, it might be possible to substitute codons for the existing amino acids in the protein genes with codons for the deficient amino acids. However, the successful implementation of this strategy may be thwarted by the following constraints. (i) The introduction of these sulphur amino acids should not perturb the molecular properties of the protein essential for its proper packaging, stability and metabolism. In certain cases, it might even be necessary to conserve the secondary structure of the mRNA itself since there is some evidence that specific sequences may be important for the stability and metabolism of *Glycine max* seed mRNAs (Schuler *et al.*,1982a). (ii) The developing seed and the plant as a whole must be able to accommodate the increased demands for sulphur amino acids on the amino acid pool if the new protein is to be efficiently synthesised. (iii) Since the storage proteins are encoded by multigene families, it will be necessary to modify all the individual genes, or at least those which are most actively transcribed. (iv) The ability to regulate gene expression, both spatially and temporally, in synchrony with the normal developmental pattern or in any other way desired will be an important objective, but also an elusive one given that so little is at present understood about the mechanisms of gene regulation in eukoryotes.

Other strategies for seed protein improvement which obviate the need for remodelling the structure of existing storage proteins have been proposed. For example, the expression of genes coding for proteins which normally occur in small amounts in the seed, but which are nutritionally more balanced, might be enhanced so that these proteins are accumulated to higher levels. Even within a particular class of storage proteins, say pea legumin, certain subunit pairs may contain reasonably high levels of sulphur amino acids while other subunit pairs contain relatively little (Casey and Short, 1981; Lycett *et al.*, 1984a). Amplification of genes coding for high-sulphur subunits, probably coupled with the silencing of low-sulphur protein genes might be a possibility. Alternatively, seed proteins might be made more nutritious by deletion of genes coding for antimetabolic or toxic proteins which contribute to the poor digestibility and low nutritional status of legume proteins.

Among the various strategies discussed above, it is widely believed that the approach likely to yield positive results the soonest involves the isolation of particular storage protein genes, the alteration of their coding sequences to correct amino acid deficiencies, and the reinsertion of the modified genes into the same or closely related species. The constraints imposed on that approach are by no means trivial but already, available techniques in recombinant DNA methodology point the way to as how they may be tackled.

X-ray diffraction techniques are the only presently available methods for determining the structures of proteins at the level of detail which will be required for protein engineering. However, protein crystallography is a laborious and time-consuming process, and the methodology for predicting three-dimensional protein structures from amino acid sequences is continually being developed (Ulmer, 1983). Some progress has already been made; for example, using a combination of sequence and physical data and computer modelling techniques, Argos *et al.* (1982) have formulated a model for the tertiary structure of zein proteins. It is anticipated that the ability to predict tertiary structures solely on the basis of primary structural data will be very important for the long-term success of protein engineering. It will enable the effects of amino acid changes on the structure of proteins to be predicted and by comparing the structures of homologous polypeptides, information will be obtained regarding which regions might tolerate amino acid substitutes without violating the structural features of the proteins. Cloning and sequencing of the genes provide the simplest and quickest means of determining the primary structures of large numbers of proteins, and the employment of these techniques therefore constitutes the first step in any project aimed at protein engineering. The individual cloned genes can also be used to assay the levels of their corresponding mRNAs and thus determine the relative efficiencies of transcription of different chromosomal genes. This information will be useful in identifying active genes whose sequences might be modifiable to the greatest effect. Once amino acid substitutions have been successfully engineered in a chosen gene, the predicted effects of these alterations on the structural and functional properties of the encoded polypeptide can be directly investigated on samples of the modified product synthesised in bacteria or yeasts before the modified gene is introduced into a plant. These approaches, when combined with the expected advances in plant transformation technology and an increasing understanding of the mechanisms of plant gene regulation, are likely to usher in a new era in the improvement of seed storage proteins by genetic manipulation.

1.5. Objectives, Rationale and Content of the Present Research.

It has already been seen that the pea seed storage proteins are encoded by multigene families. Whereas the purification and amino acid sequencing of individual gene products from these complex mixtures of homologous proteins are prohibitively difficult and time-consuming, the use of molecular cloning techniques enables the ready isolation of individual genes of absolute purity, and gives a truer picture of the complexity of the protein families. DNA sequencing, from which the protein sequence can be deduced, is simpler and more reliable than direct protein sequencing. Other important advantages accrue from studying storage protein biochemistry at the DNA level. Analyses of gene sequences reveal the nature of the primary polypeptide products synthesised which might give some insight into the post-translational processing and transport pathways of the proteins. As mentioned earlier, the specificity of storage protein synthesis constitutes a good model system for studying gene expression in eukaryotes, and the structure and organisation of storage protein genes might give clues as to how the expression of these genes is developmentally regulated. As noted in the preceding section, the cloning and structural characterisation of storage protein genes is virtually a pre-requisite for the eventual improvement of these proteins by genetic manipulation techniques, and of course, it is at the gene level that the engineering of proteins will be effected (see Ulmer, 1983).

Prior to the commencement of this research project, very little sequence data of pea storage proteins and their genes had been published. The data were limited to the N-terminal amino acid sequences of the acidic and basic subunits of legumin (Casey *et al.*, 1981a; Casey *et al.*,1981b) and the nucleotide sequences of two legumin cDNA clones which covered only ∿38% of the legumin mRNA (Croy *et al.*,1982). The cloning of a number of vicilin cDNAs had been reported (Croy *et al.*, 1982) but their sequences had not been determined. Thus, there was a need for the construction of longer legumin cDNAs and more extensive characterisation of the cloned vicilin genes.

This thesis describes the construction of a pea cDNA library and the isolation of cDNAs transcribed from legumin and vicilin mRNAs. Advances in recombinant DNA methodology have simplified the construction and screening of libraries of genomic DNA in bacteriophage λ or cosmid vectors (reviewed by Maniatis *et al.*1982; Brammar, 1982; van Embden, 1983), and the analysis of genomic clones is usually an integral part of any study of gene structure and function. However, there is often justification for the construction of cDNA clones in preference to, or in conjunction with, genomic clones. cDNA libraries are generally easier to screen than genomic libraries (see Williams, 1981), and in fact, screening of the latter frequently relies on the availability of characterised cDNA probes. This factor is particularly pertinent in the initial cloning of the storage protein genes since there are a relatively small number of chromosomal genes, whereas the mRNA transcripts encoding the major storage proteins comprise a large proportion of the total, cotyledon mRNA population (Morton *et al.*, 1983).

cDNA cloning has other advantages. It allows the determination of the sequence organisation of a gene, i.e. the precise location of its introns and of the 5' and 3' termini of its mRNA, by a comparison of the cDNA and genomic DNA nucleotide sequences. Also, if the expression of cloned genes in bacteria is desired, it is essential that the coding sequences are not interrupted by introns which are a common feature of eukaryotic genes but are absent from cytoplasmic mRNA transcripts from which cDNAs are copied (see Williams, 1981).

cDNA clones isolated from a clone bank may be analysed by a variety of techniques : hybridisation to previously characterised DNA molecules, hybrid-selection of mRNAs followed by release of the mRNA and *in vitro* translation, sizing of cDNA inserts on agarose or polyacrylamide gels, and mapping of restriction enzyme cleavage sites (see Maniatis *et al.*, 1982). Ultimately, sequence analysis must be undertaken for the fullest characterisation of a cloned gene, and the development of rapid DNA sequencing techniques has become a cornerstone of recombinant DNA technology (for reviews see Maxam and Gilbert, 1980; Messing, 1983; Davies, 1982). These techniques make it possible to determine the exact nucleotide sequences of genes and their putative controlling elements, and in this research project, the legumin and vicilin clones isolated from the cDNA library were extensively characterised, several of them to the level of DNA sequence analysis.

Although the storage proteins are synthesised in large quantities in the developing seed, the deposited protein fractions comprise mixtures of homolgous polypeptides from which it is difficult to purify individual products. Moreover, some of the subunit precursors are subject to rapid, proteolytic processing *in vivo*, and thus cannot be readily isolated. However, it may be possible to obtain useful amounts of these proteins by the expression of cloned genes in bacteria. The final part of this work describes the expression of a number of essentially full-length vicilin cDNAs in *E.coli* under the control of the bacteriophage λP_L promoter. The rationale for these expression experiments was three-fold : (i) to try and establish a general model system for studying the expression of plant genes in *E.coli*; (ii) to obtain sufficient quantities of pure ∿50000-*Mr* vicilin subunits to enable detailed investigations into the *in vivo* endoproteolytic processing of susceptible precursors; and (iii) to have the means of studying the structural and functional effects of sequence modifications introduced *in vitro* into the vicilin genes.

2. MATERIALS AND METHODS.

2.1. Materials.

2.1.1. Chemical and Biological Reagents.

Reagents, unless otherwise indicated, were obtained from BDH Chemicals Ltd., Poole, Dorset, UK, and were of analytical grade or the best available. The following materials were purchased from the designated sources.

Ethidium bromide, 4-chloro-1-napthol, adenosine triphosphate (ATP), spermidine, bovine albumin (98-99% albumin), dithiothreitol (DTT), HEPES (N-2-hydroxyethylpiperazine-N-2-ethanesulfonic acid), Tris (hydroxymethyl) aminomethane (Trizma base,reagent grade), ampicillin, kanamycin, chloramphenicol, tetracycline, RNase-A , egg white lysozyme, *E.coli* tRNA (type XXI), polyadenylic acid and herring sperm DNA were from Sigma Chemical Co., Poöle, Dorset, UK.

Caesium chloride and sodium chloride (A.R.) were from Koch-Light Ltd., Haverhill, Suffolk, UK.

Sephadex G-50, Sepharose 6B-CL and Ficoll 400 were from Pharmacia Fine Chemicals, Uppsala, Sweden.

Nitrocellulose filters (BA85, 0.45 µm) were from Schleicher and Schüll, Anderman and Co. Ltd., Kingston-upon-Thames, Surrey, UK.

3MM paper and DEAE-cellulose (DE-81) paper were from Whatman Ltd., Maidstone, Kent, UK.

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

BBL trypticase peptone was from Becton Dickinson and Co., Cockeysville, MD, USA.

Oligo $(dT)_{12-18}$ was from Collaborative Research Inc., Waltham MA, USA.

Restriction endonucleases were from Bethesda Research Laboratories (UK) Ltd., Cambridge, UK, The Boehringer Corporation (London)Ltd., Lewes, East Sussex, U.K., or New England Biolabs, CP Laboratories Ltd., Bishop's Stortford, Herts., UK.

Calf intestinal alkaline phosphatase, endonuclease-free *E.coli* DNA polymerase 1, T4 polynucleotide kinase, T4 DNA ligase and S1 nuclease were from The Boehringer Corporation (London) Ltd.

E.coli DNA polymerase 1 large fragment (Klenow enzyme), T4 DNA polymerase, BamHI linkers (decameric) and agarose (electrophoresis grade) were from Bethesda Research Laboratories (UK) Ltd.

Mung-bean nuclease, deoxy- and dideoxynucleoside triphosphates were from Pharmacia P.L. Biochemicals Inc., Pharmacia (Great Britain) Ltd., Milton Keynes, Bucks., UK.

DNase 1 (DPFF) was from Worthington Biochemicals, Millipore (UK) Ltd., London, UK.

Avian myeloblastosis virus (AMV) reverse transcriptase was from the Division of Cancer Cause and Prevention, National Cancer Institute, NIH, Bethesda, MD, USA.

Placental ribonuclease inhibitor (RNasin) was from Biotech, Madison, Wisconsin, USA.

Radiochemicals and nick translation kits were from Amersham International plc, Amersham, Bucks., UK.

Poly(A)$^+$ mRNA, prepared from 14-day-old cotyledons of *Pisum sativum* L. var. Feltham First (Sutton Seeds Ltd., Reading, Berks., UK), was a generous gift from Dr. I.M. Evans.

Genomic DNA prepared from leaves of *Pisum sativum* L. var. Feltham First, and affinity-purified, rabbit antivicilin IgG were supplied by Dr. J.A. Gatehouse and Mr.D. Bown.

2.1.2. Bacterial Strains and Plasmids.

All bacterial strains used in this work were derivatives of

E.coli K-12 and are listed in Table 1. Plasmids used in cloning experiments and as sources of DNA fragments are also listed in Table 1.

TABLE 1. Properties of *E.coli* Strains and Plasmids Used.

Bacterial Strains	Genetic Characters	Reference/Source
910	Ap^S Tc^S(803 *SupE SupF* $RecBC^-$).	W.J.Brammar,Dept. of Biochemistry, University of Leicester, UK.
K-12ΔH1Δ*trp*	Sm^R,*lacZ*am,Δ*bio-uvrB*,Δ*trpEA*2. (λ*N*am7,*N*am53,*c*I857,ΔH1)	Remaut *et al.*,1981
SG4044 [p*c*I857]	lac^-,*lon*Δ100,Δ(*gal-blu*)*strA* *c*I857,km^R	Remaut *et al.*,1983a; Remaut *et al.*,1983b.
N99λcI^+	$lacZ^-$, $galK^-$, thi^-,su^o λcI^+	Young *et.al.*; 1983.
N99λ*c*I857	$lacZ^-$, $galK^-$,thi^-,su^o,λ*c*I857	Rosenberg *et al.*,1983.
N5151(*c*I*ts*857)	λ*c*I857	Young *et al.*, 1983.
Plasmids		
pBR322	Ap^R, Tc^R	Bolivar *et al.*,1977
pDUB2	Ap^R, 50K vic^+	Lycett *et al.*,1983a.
pDUB3	Ap^R, leg^+	Croy *et al.*, 1982.
pDUB4	Ap^R, 47K vic^+	Lycett *et al.*, 1983a.
pPLc24	Ap^R, $\lambda O_L P_L{}^+$	Remaut *et al.*, 1981.
pPLc245	Ap^R, $\lambda O_L P_L{}^+$	Remaut *et al.*, 1983a.
pAS1	Ap^R, $\lambda O_L P_L{}^+$	Rosenberg *et al.*,1983.

Key: [p*c*I857] , harbouring plasmid p*c*I857; Sm^R, resistance to streptomycin; km^R, resistance to kanamycin; Ap^R, resistance to ampicillin; Tc^R, resistance to tetracyline; 50K vic^+,presence of cDNA coding for part of pea vicilin 50000-*M*r subunit; 47K vic^+, presence of cDNA coding for part of pea vicilin 47000-*M*r subunit; leg^+, presence of cDNA coding for part of pea legumin subunit; $\lambda O_L P_L{}^+$, presence of the λleftward operator-promoter region;

2.1.3. Notes on the *E.coli* Expression Systems.

The expression plasmid, pPLc24, contains the leftward operator-promoter region ($\lambda O_L P_L$) of bacteriophage λ, followed by the translation initiation signals and the N-terminal region of the bacteriophage MS2 replicase gene cloned into a pBR322 derivative (Remaut *et al.*, 1981; Table 1). Insertion of foreign gene sequences in the correct reading frame at a unique BamHI site in the plasmid leads to the synthesis of fusion proteins containing the N-terminal 98 amino acid residues of MS2 replicase.

The plasmid, pPLc245, was designed for the expression of unfused proteins and is a derivative of pPLc24 in which a polylinker sequence has been inserted immediately downstream from the initiation ATG codon of the MS2 replicase gene (Remaut *et al.*,1983a;Table 1). The G-residue of the initiation codon constitutes the 3' end of a unique SalI cleavage site within the polylinker sequence. Thus linearisation of pPLc245, followed by removal of the protruding 5' terminus to give a blunt end, leaves the ATG codon accessible for direct coupling to the coding sequence of a foreign gene. Genes which contain compatible restriction enzyme sites near their N-termini can also be ligated to pPLc245 via the cohesive ends of the linearised vector.

The plasmid pAS1, like pPLc245, can be used for expression of unfused proteins. The vector is a pBR322 derivative into which has been cloned the $\lambda O_L P_L$ region and translation initiation signals from the λ*c*II gene (Rosenberg *et al.*,1983; Table 1). The G-residue of the initiation ATG of the λ*c*II gene constitutes the 3' end of a unique BamHI cleavage site which is exactly analagous to the SalI site in pPLc245. Genes which contain compatible restriction enzyme sites near their N-termini can also be inserted directly into the BamHI site of pAS1.

In all three expression plasmids above, control over gene expression is effected by maintaining the plasmid in a defective lysogen carrying a temperature-sensitive mutation (λ*c*I*ts*857) in the λ*c*I gene (see Table 1 for examples of such strains). A functional *c*I repressor is synthesised at low temperatures (30°C) but the repressor is inactivated at elevated temperatures (42°C); thus cells harbouring the expression plasmid can be grown to high density at 30°C without expression of the inserted gene, and subsequently induced to synthesize the required product by switching the culture to 42°C.

2.2. Methods.

2.2.1. Biochemical Techniques.

2.2.1.1. Glassware and Plasticware.

All plasticware used for handling nucleic acid samples was sterilised by autoclaving before use. All glassware and plastic eppendorf tubes were siliconised with "Repelcote" (Hopkins and Williams, Romford, U.K.).prior to being autoclaved.

2.2.1.2. Alcohol Precipitation of DNA.

0.1 volume of 3M sodium acetate pH 5.2 and 2.5 volumes of ethanol were added to the DNA solution and kept at -80°C for 20 min. or at -20°C overnight. The precipitated DNA was pelleted at 12000g for 15 min. (MSE Micro Centaur microcentrifuge) for small samples, or at 25000g for 30 min. (Sorvall RC-5B centrifuge) for larger samples. The pellet was usually washed twice in 70% (v/v) ethanol, dried briefly under vacuuum, and redissolved in a small volume of water or TE buffer (10mM Tris-HCl pH7.5, 1mM EDTA). To minimize the volume of the sample to be centrifuged, isopropanol was sometimes used instead of ethanol. In these cases, 0.6 - 1.0 volumes of isopropanol were added to the DNA solution and the mixture was kept at -20°C for 20 min. prior to centrifugation.

2.2.1.3. Phenol Extraction of DNA Samples.

Solutions of DNA were deproteinised by two successive extractions with phenol-chloroform-isoamyl alcohol (25:24:1 v/v)—henceforth referred to simply as "phenol". 1.5 volumes of phenol were added to the DNA sample and mixed by vortexing. The aqueous and phenolic phases were separated by a brief centrifugation (∿15s in a microcentrifuge). The upper aqueous phase was transferred to a fresh tube and the phenol extraction was repeated. When extracting minute amounts of valuable DNA, the phenol phases were "back-extracted" with equal volumes of TE buffer and the aqueous phase from a back-extraction was combined with the original aqueous phase. Phenol extractions were followed by two to three extractions with 3 volumes of diethyl ether to remove the remaining traces of phenol. The DNA was recovered by alcohol precipitation.

2.2.1.4 Dialysis of DNA Solutions.

Suitable lengths of visking dialysis tubing (Size 1-8/32"; Medicell International Ltd., London, UK) were boiled for 20 min in 10 mM EDTA, then thoroughly rinsed in distilled water. One end of the tubing was closed with a knot and the DNA sample was pipetted in through the open end. A space was left above the solution to allow for an increase in the liquid volume, and the open end of the tubing was knotted. The sealed dialysis bag was then placed in a large volume of TE buffer ($>$1l) which was stirred for several hours at 4°C. The TE buffer was changed 2-3 times over a period of $\sim$24 hr.

2.2.1.5. Spectrophotometric Quantitation of DNA Solutions.

The optical density (O.D.) of DNA solutions in quartz glass cells were recorded from 320 to 230 nm in a Pye Unicam SP8-150 uv/vis spectrophotometer operated in the scanning mode.

An O.D.$_{260}$ of 0.02 corresponds to a DNA concentration of $\sim$1μg/ml. A pure DNA sample has an O.D.$_{260}$/O.D.$_{280}$ ratio of $\sim$1.8 and the O.D.$_{260}$/O.D.$_{235}$ ratio is higher than the O.D.$_{260}$/O.D.$_{280}$ ratio. Also, the O.D.$_{320}$ is zero. Deviations from these relationships indicated the presence of protein, phenol or particulate contaminants, in which cases accurate quantitation of the DNA was not possible.

2.2.1.6. Storage of Bacteria.

Bacterial colonies were regularly stored at 4°C for up to 6 weeks on inverted agar plates sealed with Nescofilm (Nippon Shoji Kaisha Ltd., Osaka, Japan). For long-term storage, bacterial lawns grown from single colonies on selective agar plates were transferred to sterile 2 ml aliquots of a solution comprising 50% L broth and 40% glycerol, mixed thoroughly by vortexing, and stored at -80°C.

2.2.2. Rapid Mini-preparation of Plasmid DNA.

The method used was essentially that of Birnboim and Doly (1979) with minor modifications as described below. The plasmid-bearing strain was grown to saturation at 37°C (or 30°C for temperature-inducible expression plasmids) in 10 ml of L broth supplemented with appropriate antibioties. The cells were pelleted by centrifugation at ∿6000g for 5 min in an MSE bench centrifuge (using the culture bottles as centrifuge tubes), and resuspended by vortexing in 200 μℓ of freshly prepared 50mM glucose, 25 mM Tris-HCl pH 8.0, 10 mM EDTA pH8.0, 4 mg/ml lysozyme. The suspension was transferred to a 1.5 ml-eppendorf tube and placed on ice for 15-20 mins. 500 μℓ of freshly prepared 0.2M NaOH, 1% SDS were added, mixed gently by inversion, and kept on ice for 5 min. 375 μℓ of 3M sodium acetate pH 4.8 were added and thoroughly mixed. The mixture was placed on ice for 30 min. with vigorous agitation every 5 min. during that period. The sample was centrifuged at 12000g for 15 min. Cold isopropanol (0.6ml) was added to the supernatant (1.0 ml), mixed by inversion, and kept at -20°C for 15 min. The DNA was pelleted at 12000g for 10 min., resuspended in 400 μℓ of TE buffer, and reprecipitated with ethanol. The precipitate was again pelleted, washed twice in 70% ethanol, and dried under vacuum. The DNA pellet was dissolved in 20-40 μℓ of TE buffer and stored at -80°C.

2.2.3. Large-scale Preparation of Plasmid DNA.

Two methods were routinely used to prepare plasmid DNA on a large scale. One was adopted from the procedures of Clewell(1972) and Katz *et al.*(1977) and involved chloramphenicol amplification of the plasmid followed by lysis of the bacteria with SDS. The second procedure was essentially a scaled-up version of the alkaline-lysis mini-prep method described previously (section 2.2.2.). It was used for preparing λP_L-containing expression plasmids since Bernard *et al.*(1979) had warned against their chloramphenicol amplification on the basis that continued plasmid replication in the absence of protein synthesis may result in the activation of the P_L promoter.

2.2.3.1. SDS-Lysis Method. .

The plasmid-bearing strain was grown at 37°C in 250 ml of L broth containing the appropriate antibiotics to an $O.D._{650}$ of 0.8. Chloramphenicol (170 μg/ml) was added and incubation was continued at 37°C for 15-20 hr. The cells were harvested by centrifugation at 6000g for 10 min. at 4°C, resuspended in 5.0 ml of 25% sucrose in 50mM Tris-HCl pH 8.0, and chilled on ice. 1.0 ml of a freshly prepared lysozyme solution (10 mg/ml in 25% sucrose, 50 mM Tris-HCl pH 8.0) was added and incubated with shaking for 2 min at 37°C, then for an additional 10 min on ice. 5.0 ml of 0.2M EDTA were added, and the shaking on ice was continued for 10 min. 1.0 ml of 20% SDS was added and the mixture was rocked gently at room temperature until the suspension clarified. 3.0 ml of 5M NaCl were added, mixed thoroughly, and stored on ice for at least 2 hr. The suspension was centrifuged at 27000g for 90 min at 4°C. 0.6 ml of a 10 mg/ml stock of ethidium bromide (EtdBr) was added to the supernatant, followed by the addition of CsCl to 48.4%(w/w). The solution was stored on ice for 30-60 min and then centrifuged at 12000g for 30 min at 4°C. The red pellicle on the surface of the supernatant was removed and the supernatant was centrifuged in a Beckman vertical rotor VTi50 at 44000 rpm for 18-24 hr at 15°C. The lower plasmid band was removed with a syringe and needle inserted through the side of the centrifuge tube. The harvested plasmid DNA was sometimes repurified by centrifugation through a second CsCl gradient as described above. The EtdBr was extracted 4 or 5 times with CsCl-saturated isopropanol and the plasmid was dialysed overnight against TE buffer. The DNA was then precipitated with ethanol, washed twice with 70% ethanol and redissolved in 200-300 μℓ of TE buffer.

2.2.3.2. Alkaline Lysis Method.

The plasmid-bearing strain was grown to saturation at 30°C in 250 ml of antibiotic-supplemented L broth. The cells were harvested by centrifugation at 6000g for 10 min at 4°C, resuspended in 2.0 ml of 50mM glucose, 10mM EDTA, 25mM Tris-HCl pH 8.0, placed on ice, and 4.0ml of a freshly prepared lysozyme solution (4mg/ml in 50mM glucose, 10mM EDTA, 25 mM Tris-HCl pH 8.0) were added. The mixture was incubated with shaking for 2 min at 37°C, then for an additional 20 min on ice. The suspension was transferred to a 30 ml-Corex tube. 12ml of 0.2M NaOH, 1% SDS were added and mixed until nearly homogeneous. The mixture was stored on ice for 10 min. 9.0 ml of ice-cold 3M sodium acetate pH 4.8 were added and thoroughly mixed. The mixture was kept on ice for 45 min with occasional inversions. The precipitate was pelleted at 12000g for 30 min at 4°C. An approximately equal volume of isopropanol (27ml) was added to the supernatant, mixed, and stored at -20°C for 15 min. The DNA was pelleted by centrifugation, washed once in 70% ethanol, and dried briefly in a vacuum dessicator. The pellet was redissolved in TE buffer, followed by the addition of EtdBr to 400 µg/ml and CsCl to 48.4% (w/w). Purification of the plasmid through one or two EtdBr-CsCl gradients was carried out as described in the preceding section.

2.2.4. Enzymic Reactions Used Routinely in DNA Manipulations.

2.2.4.1. Restriction Endonuclease Digestion.

DNA molecules were digested with type-II restriction endonucleases in one of the 4 buffers recommended by Maniatis *et al.* (1982). The buffers, modified to include spermidine, were those shown in Table 2.

TABLE 2. Restriction Endonuclease Digestion Buffers.

Buffer	Components (mM)					
	Tris-HCl pH7.5	$MgCl_2$	DTT	Spermidine	NaCl	KCl
Low Salt	10	10	1.0	2.0	-	-
Medium Salt	10	10	1.0	2.0	50	-
High Salt	50	10	1.0	2.0	100	-
SmaI	10(pH8.0)	10	1.0	2.0	-	20

Generally, the enzymes were used at a concentration of 2-5u/µg of DNA and incubated at the temperature recommended by the manufacturers for 1-3hr. Many of the enzymes have been shown to work adequately at different NaCl concentrations (New England Biolabs 1983/84 Catalogue); thus, multiple digestions could usually be performed simultaneously in the same buffer. For digestion of mini-prep plasmid DNA, 25µg/ml of RNase (pre-boiled for 30 min to inactivate contaminating DNases) were included in the reaction mixture.

2.2.4.2. 5'-Dephosphorylation of DNA with Alkaline Phosphatase.

The 5' phosphate groups of DNA molecules were removed by treatment with calf intestine alkaline phosphatase in 50mM Tris-HCl pH 9.0, 1mM $MgCl_2$, 0.1mM $ZnCl_2$ and 1mM spermidine (Maniatis *et al.*, 1982). For fragments with protruding 5' termini, the reaction mixture was incubated for 1hr at 37°C with 0.2u/µg of DNA. To dephosphorylate blunt-ended molecules, the reaction was incubated for 15 min periods first at 37°C, then at 56°C. A second aliquot of phosphatase was then added and the incubations at both temperatures repeated. Following the phosphatase reaction, the enzyme was removed by two phenol extractions.

2.2.4.3. DNA Ligation.

ds-DNA molecules with compatible, protruding ends or blunt ends were covalently joined by treatment with T4 DNA ligase in a minimal volume of KLP buffer (50mM Tris-HCl pH7.5, 10mM $MgCl_2$, 10mM DTT - so designated because the same buffer was used for *k*inase, *l*igase and *p*olymerase reactions (Sippel *et al.*,1978)) containing 1mM ATP. Cohesive termini were ligated at 12°C for 12-20 hr whereas blunt ends were ligated at 6-8°C for up to 48 hr.

2.2.4.4. 3'→5' Exonuclease Digestion of ds-DNA with T4 DNA Polymerase.

The 3'-termini of ds-DNA fragments were progressively digested with T4 DNA polymerase (0.6 u/μg DNA) in 100 μℓ of 33mM Tris-acetate pH 7.9, 66mM potassium acetate, 10mM magnesium acetate, 0.5 mM DTT, 0.1 mg/ml BSA at 37°C. Under these conditions, the rate of exonuclease excision from each 3' end is ∿10 nucleotides/min (Maniatis *et al.*,1982). Digestion was terminated at a selected nucleotide by including the appropriate dNTP (200μM) in the reaction mixture such that digestion proceeded until a nucleotide complementary to that dNTP was exposed on the opposite DNA strand. When that nucleotide was exposed, the 5'→3' polymerase activity of the enzyme blocked any further exonuclease activity. The above buffer was used both for cleavage of DNA with BamHI and for subsequent exonuclease digestion when the restriction reaction immediately preceeded the T4 polymerase reaction.

2.2.4.5. Digestion of ss-DNA with Mung-bean Nuclease.

Single-stranded protruding termini on ds-DNA molecules were removed by treatment with mung-bean muclease (5μ/μg DNA) in 50 mM sodium acetate pH5.2, 50mM NaCl, 2mM $ZnCl_2$ 1mM DTT for 20 min at 22°C (Kroeker *et al.*,1976).

2.2.5. Agarose Gel Electrophoresis.

DNA fragments in the size range 0.1-30kb were resolved on agarose gels of various concentrations as indicated in Table 3.

Table 3. Applicability of Agarose Gels of Various Concentrations for Fractionation of DNA Fragments.

Agarose Concentration (%)	Approx. size-range of efficiently resolved linear DNA fragments (Kb).
0.5	1.0 - 30
0.8	0.6 - 10
1.0	0.4 - 6
1.5	0.2 - 4
2.0	0.1 - 3

Horizontal gels, submerged in electrophoresis buffer (40mM Tris-acetate pH7.7, 2mM EDTA, 1μg/ml EtdBr) were used. Gels of the appropriate concentration measuring 18.5 x 15.2 x 0.6 cm, were prepared as described by Maniatis *et al.*(1982) except that a perspex gel mould (Shandon Southern Products Ltd., Cheshire, U.K.), held in place on a horizontal glass plate by a thin layer of vacuum grease, was used and EtdBr was added to the gel solution to a final concentration of 1.0μg/ml. DNA samples were mixed with 0.3 volumes of agarose beads (10mM Tris-HCl pH 8.0, 10mM EDTA, 30%(v/v) glycerol, 0.1%(w/v) bromophenol blue, 0.1%(w/v) xylene cyanol, 0.2%(w/v) agarose - autoclaved, then extruded through a syringe and fine needle when set) loaded into 0.9-1.2 cm wide slots and electrophoresed at 1.6V/cm. DNA bands in the gel were visualised by UV light (254 nm)- induced EtdBr fluorescence. Gels were photographed through a red-orange filter (Kodak 23A Wratten) using transmitted UV light at 254nm and Polaroid Type 667 (3000 ASA) film. An exposure of 10s at f5.6 enabled as little as ∿8ng of DNA to be detected.

2.2.6. Recovery of DNA from Agarose Gels.

The method of Dretzen *et al.*(1981) was used with minor modifications. Strips of DEAE-cellulose paper (Whatman DE81) were processed by soaking for several hours in 2.5M NaCl, washed thoroughly with water, and stored dry between sheets of 3MM paper at room temperature. After gel electrophoresis, strips of the DEAE-cellulose paper were inserted into slits made immediately infront and behind the desired fragment. Electrophoresis was resumed until the fragment had completely entered the paper. The strip of paper inserted behind the band served to prevent contamination by larger fragments and was subsequently discarded. The DEAE-cellulose paper containing the desired fragment was washed in distilled water, and blotted dry on 3MM paper. The immobilised DNA was located on the paper by UV fluorescence and the excess paper was trimmed off. The paper was then placed in a 1.5ml-eppendorf tube and 300μl of elution buffer (1.5M NaCl, 20mM Tris-HCl pH 7.5, 1mM EDTA) per 30mm^2 of paper were added. The paper was shredded by vortexing and was incubated at 37^{o}C for 2hr with occasional agitation. The slurry was then transferred to a 1ml pipette-tip plugged with siliconized glass-wool and the eluate was "blown out" into an eppendorf tube using a stream of pressurised nitrogen gas. The shredded paper was washed twice with 100 μℓ aliquots of elution buffer and the washings were combined with the primary eluate. The total eluate was centrifuged at 12000g for 3min. and the supernatant was transferred to a fresh tube. It was then extracted with 2 volumes of elution buffer-saturated isoamyl alcohol, and the DNA recovered by ethanol precipitation. DNA fragments recovered by this procedure required no further purification before subsequent enzymic reactions. Recovery was estimated to be 70-80% for linear molecules of 0.1-6.0 kb.

2.2.7. Fractionation of DNA on Polyacrylamide Gels.

Polyacrylamide slab gels were used to (i) analyse, oligomeric forms of BamHI linkers (section 2.2.10.3); (ii) isolate ^{32}P-labelled DNA fragments from preparative gels (section 2.2.20.1); and (iii) obtain high resolution of ss-DNA molecules for DNA sequencing (section 2.2.20.3).

2.2.7.1. Fractionation of ^{32}P-labelled, Oligomeric Linkers.

10% gels 15 cm long x 18 cm wide x 0.15cm thick, were prepared using the recipe in Table 4 and run in a Studier-type electrophoresis apparatus (Studier, 1973) obtained from Raven Scientific Ltd., Haverhill, U.K. Samples containing 0.3 volumes of a glycerol dye solution (10mM Tris-HCl pH8.0, 10mM EDTA, 80%(v/v) glycerol, 0.1%(v/v) bromophenol blue, 0.1%(w/v) xylene cyanol) were electrophoresed at 5V/cm until the bromophenol blue was ∿5cm from the bottom of the gel.

Table 4. Recipes for the Preparation of Different Polyacrylamide Gel Types.

	Final acrylamide concentration (%)			
	10[(a)]	5[(b)]	6[(c)]	8[(c)]
Ingredients	Volume and weight required.			
40% acrylamide stock solution (ml)[(d)]	20.0	20.0	7.5	10.0
10 x TBE buffer (ml)[(e)]	8.0	16.0	5.0	5.0
Urea (g)	-	-	24.0g	24.0g
Glycerol (ml)	20.0	-	-	-
Water to final volume (ml)	80	160	50	50
	Mix and deaerate under vacuum			
20% (w/v) Ammonium persulphate (ml)	0.8	1.1	0.3	0.3
TEMED (ml)	0.025	0.1	0.02	0.02
	Mix and pour gel immediately			

a. Gels used for the analysis of oligomeric BamHI linkers (see section 2.2.7.1).

b. Preparative gels for isolation of DNA fragments for DNA sequencing (section 2.2.7.2)

c. DNA sequencing gels (section 2.2.7.3)

d. 40% acrylamide stock : 38% (w/v) acrylamide, 2% (w/v) bisacrylamide.

e. 10x TBE buffer : 108g Tris base, 55.0g boric acid, 9.3g EDTA-$Na_2.2H_2O$ per litre (pH∿8.3). 1X buffer was used as the electrophoresis buffer.

2.2.7.2. Preparative Gel Electrophoresis.

5% gels, 36cm long x 18 cm wide x 0.15cm thick, were prepared according to Table 4 (apparatus described by Davies, 1982). 0.5 volumes of a glycerol dye solution (see preceding section) were added to the DNA samples and gels were run at ∿14V/cm until the migration of the marker dyes indicated adequate resolution of the DNA fragments : the bromophenol blue comigrated with fragments of ∿40bp while the xylene cyanol comigrated with fragments of ∿190bp.

2.2.7.3. DNA Sequencing Gels.

Denaturing gels, 38 x 18 x 0.035cm, of 6 or 8% polyacrylamide containing 8M urea (see Table 4) were used for electrophoresis of sequencing samples. The gels were constructed and run essentially as described by Davies (1982). Electrophoresis was carried out at ∿25mA (1500-1700V) which maintained the temperature of the gel at ∿70°C. Depending on the length of the fragment being sequenced, multiple sample loadings (up to three) were applied to each gel. The intervals between each loading were judged by the migration of the marker dyes, and were chosen so as to allow at least 20 nucleotides of sequence overlap between successive loadings. The bromophenol blue comigrated with single-stranded fragments of ∿23 and ∿19.nucleotides long on 6 and 8% gels respectively, while the xylene cyanol comigrated with fragments of ∿98 and ∿72 nucleotides long.

2.2.8. Fractionation of Denatured Proteins on SDS-Polyacrylamide Gels.

Mixtures of polypeptides, dissolved and denatured by boiling in SDS sample buffer (section 2.2.22.) were fractionated on 12.5, 15 or 17% SDS-polyacrylamide slab gels using a discontinuous buffer system (Laemmli, 1970). Recipes for the preparation of the resolving and stacking gels are given in Table 5. Gels, 15 x 18 x 0.15cm were constructed and run in a Studier - type gel apparatus essentially as described by Hames (1981). The reservoir buffer comprised 192 mM glycine, 25mM Tris base, 0.1% SDS. Three drops of tracking dye (1%(w/v) bromophenol blue in ethanol) were added to the buffer in the top reservoir prior to the start of electrophoresis and the gels were run at 8mA overnight or at 25mA for ∿4.5hr until the bromophenol blue reached the bottom of the gel. On completion of electrophoresis, the proteins were visualised either by staining the gel, by fluorography if tritium labelled (section 2.2.17.), or by electroblotting onto nitrocellulose paper followed by immunological screening (section 2.2.25). The gel was stained by soaking for several hours in ∿350ml of Kenacid blue stain (0.05%(w/v) Kenacid blue R in 50% (v/v) methanol, 7%(v/v) acetic acid). Excess stain was removed by soaking the gel in 2-3 changes of destain solution (50% (v/v) methanol, 7%(v/v) acetic acid) over a period of ∿8hr.

Table 5. Recipes for the preparation of SDS-Polyacrylamide gels using the discontinuous buffer system.

Components	Final acrylamide concentration (%)			
	12.5(a)	15(a)	17(a)	3(b)
	Volume (ml)			
30% acrylamide stock (c)	25	30	34	2.0
1.0M Tris-HCl pH8.8(a) or 6.8(b)	22.5	22.5	22.5	2.5
H_2O to final volume	60	60	60	20
	Mix and deaerate under vacuum			
10%(w/v) SDS	0.6	0.6	0.6	0.2
1.5%(w/v) Ammonium persulphate	1.5	1.5	1.5	0.5
TEMED	0.02	0.02	0.02	0.02
	Mix and pour immediately			

a. Recipe for the resolving gel.
b. Recipe for the stacking gel.
c. 30% acrylamide stock : 30%(w/v) acrylamide, 0.135%(w/v) bisacrylamide for the resolving gel, and 30% (w/v) acrylamide, 0.433% (w/v) bisacrylamide for the stacking gel.

2.2.9. Transformation of *E.coli* Cells by Plasmid DNA.

E.coli cells were made competent for DNA transformation by the procedure of Dagert and Ehrlich (1979). Briefly, 50ml of the *E.coli* culture were grown at 37°C (or 30°C with lysogens to be transformed with temperature-inducible expression plasmids) to an $O.D._{650nm}$ of 0.2. The culture was chilled on ice for 10 min and the cells were pelleted at 6000g for 5 min at 4°C. The pellet was resuspended in 20ml of ice-cold 0.1M $CaCl_2$ and placed on ice for 20-30 min. The cells were again harvested by centrifugation, resuspended in 2ml of ice-cold 0.1M $CaCl_2$, and kept on ice until used. The maximum transformation efficiency was obtained after 24 hr on ice.

For transformation, the DNA, dissolved in 5-20 μl of water, TE or ligation buffer, was added to 100-200 μl of the competent cell suspension. The mixture was kept on ice for 20 min. and then incubated at 37°C (or 32°C with temperature-inducible expression plasmids) for 5 min. 0.8ml of L broth was added, mixed, and incubated for 1 hr at 37°C (or 30°C where appropriate) without shaking. Aliquots of the transformation mixture (10-200μℓ) were spread onto selective agar plates and incubated overnight at 37°C (or 30°C where appropriate).

2.2.10. Construction of a Pea Cotyledon cDNA Library.

2.2.10.1. Preparation of poly(A)$^+$ RNA.

Poly(A)$^+$ RNA was a gift from Dr. I.M.Evans and was prepared from polyribosomes isolated from pea cotyledons 14 days after flowering (Evans *et al.*,1979), and purified twice on oligo(dT)-cellulose columns (Evans *et al.*,1980).

2.2.10.2. Synthesis and Size-Fractionation of ds-cDNA.

The synthesis of ds-cDNA was based on the method of Wickens *et al.*(1978). For first strand cDNA synthesis, 6.0μg of poly(A)$^+$ mRNA were incubated at 37°C for 30 min in 100μℓ of 50mM Tris-HCl pH 8.3, 100mM KCl, 8mM MgCl$_2$, 8mM DTT, 0.8mM of each dNTP except dCTP, 50μCi ^{3}H-dCTP, 30 units RNasin, 0.4μg oligo(dT)$_{12-18}$ and 170 units AMV reverse transcriptase. The mixture was then heated at 100°C for 3 min and cooled rapidly on ice. The second cDNA strand was synthesised by adding to the ss-cDNA mixture an equal volume of a buffer comprising 100mM HEPES pH 6.9, 200mM KCl, 0.32mM of each dNTP, 50μCi ^{3}H-dCTP and 20 units of *E.coli* DNA polymerase 1 large fragment, and was incubated at 37°C for 1hr. The reaction mixture was then phenol extracted, and the DNA was separated from unincorporated dNTPs by chromotography on a column of Sephadex G-50 equilibrated and eluted with 300mM NaCl, 50mM Tris-HCl pH7.5. Fractions containing the cDNA as determined by Cerenkov counting were pooled, and the DNA was recovered by ethanol precipitation in the presence of 10μg of carrier *E.coli* tRNA. The cDNA was digested with 1000 units of S1 nuclease in 34 μℓ of 200mM NaCl, 1mM ZnSO$_4$, 50mM sodium acetate pH4.4, first at 37°C for 30 min and then at 25°C for an additional 30 min. The reaction was stopped by the addition of EDTA to 5mM followed by phenol extraction and ethanol precipitation. To maximize the number of molecules with perfectly blunt ends, the ds-cDNA was treated with 1 unit of *E.coli* DNA polymerase 1 for 30 min at 13°C in 20μℓ of KLP buffer (see section 2.2.4.3.) containing 0.25mM of all four dNTPs. The mixture was then electrophoresed on a 0.5% agarose gel and two fractions comprising molecules of 1-2kb and >2kb were recovered from the gel as previously described (section 2.2.6.), and each redissolved in 20µℓ of KLP buffer. The following series of reactions were based on the procedures of Sippel *et al.* (1978).

2.2.10.3. Test Phosphorylation, Ligation and Restriction of BamHI linkers.

1.35µg (200pmol) of BamHI linkers (CCGGATCCGG) were treated with 4.5u of T4 polynucleotide kinase in 12µℓ of KLP buffer (see section 2.2.4.3.) containing 40µCi of γ-^{32}P-ATP. After a 30 min incubation at 25°C, cold ATP was added to 1mM and incubation was continued at 25°C for 3hr. The kinased linkers were ligated with 3u of T4 DNA ligase in 40µℓ of KLP buffer containing 1mM ATP at 12°C for 15hr. After the ligation reaction, the enzyme was inactivated by heating at 70°C for 10 min. A 10µℓ (50 pmol) aliquot was withdrawn. Its volume was increased to 50 µℓ containing 10mM Tris-HCl pH7.5, 10 mM $MgCl_2$, 10mM NaCl, 5mM DTT and varying amounts of BamHI (see Results, section 3.1.2.), and was incubated at 37°C for 2hr. The restricted sample and an equivalent amount of the ligated, kinased linkers (50pmol) were electrophoresed on a 10% polyacrylamide gel and the labelled oligonucleotides were visualised by autoradiography of the frozen gel.

2.2.10.4. Linkering and Restriction of cDNAs.

Two 400 pmol aliquots of BamHI linkers were each treated with 9u of T4 polynucleotide kinase at 25°C for 3.5hr in 12µℓ of KLP buffer containing 1mM ATP. The kinased linkers were ligated to the two cDNA size classes (1-2Kb and >2Kb) with 5u of T4 DNA ligase at 12°C for 16hr in 52µℓ of KLP buffer containing 1mM ATP. The ligase was then inactivated by heating at 70°C for 10 min. Each linkered cDNA sample was then digested with BamHI and, following the addition of 20 µg of *E.coli* tRNA, extracted with phenol. The cDNA was separated from the monomeric linkers by chromatography on a column of Sepharose 6B-CL equilibrated and eluted with 10mM Tris-HCl pH 7.5, 100mM NaCl, 1mM EDTA. Fractions containing the cDNA (as indicated by ^{3}H- counting) were pooled and ethanol-precipitated. The DNA precipitates were washed with 70% ethanol and resuspended in 40µℓ of KLP buffer. It was estimated by ^{3}H-counting that 5µg of the 1-2Kb cDNA species and 2µg of the ≥2Kb species had been recovered.

2.2.10.5. Preparation of the Plasmid Vector for Ligation to Linkered cDNAs.

30µg of pBR322 were digested with BamHI and the restricted DNA was extracted with phenol, recovered by ethanol precipitation, and redissolved in 40µℓ of water. Two strategies were adopted for ligating the linkered cDNAs to the plasmid vector. In one, the cDNA was ligated to a molar excess of the linearised plasmid, and recombinant molecules were recovered after fractionation of the ligation products by electrophoresis on an agarose gel.

In the other strategy, the linearised plasmid was first 5'-dephosphorylated and then ligated to the cDNA. The ligation products were then used directly for transformation. 20µg of the BamHI-linearised plasmid were treated with alkaline phosphatase and the reaction stopped by phenol extraction. The DNA was ethanol-precipitated and resuspended in 100µℓ of water. A small sample (∿0.2µg) was electrophoresed through an agarose gel to verify that linearization of the plasmid had gone to completion and the DNA had not been degraded by phosphatase treatment.

2.2.10.6. Ligation of cDNAs to pBR322.

i) Phosphatase-treated plasmid.

0.5µg of the 1-2Kb, BamHI-digested, linkered cDNA was ligated to a 9-fold molar excess of the BamHI-digested, 5'-dephosphorylated pBR322. In a parallel reaction, 0.5µg of the >2Kb, BamHI-digested cDNA was ligated to a 6-fold molar excess of the phosphatased vector. A small sample of each ligation mixture (≡0.5µg of plasmid) was electrophoresed on an agarose gel to monitor the products of the ligation reaction. The remainders of the ligated DNA were ethanol-precipitated and redissolved in 100µℓ aliquots of 1mM EDTA.

ii) Nonphosphatased Plasmid.

0.25µg of the 1-2Kb, BamHI-restricted cDNA was ligated to a 9-fold molar excess of BamHI-restricted pBR322, while 0.25µg of the >2Kb BamHI-cut cDNA was ligated to a 6-fold molar excess of BamHI-restricted pBR322.

2.2.10.7. Fractionation and Isolation of cDNA-plasmid Chimaeras.

The ligation products formed by the ligation of the two cDNA species to the nonphosphatased pBR322 were electrophoresed on a 0.5% agarose gel. Vector-cDNA hybrids were recovered from the gel as previously described (section 2.2.6.), and dissolved in 50µℓ aliquots of 1mM EDTA.

2.2.10.8. Transformation of *E.coli* and Screening for Tetracycline-sensitive (Tc^s) Transformants.

Competent *E.coli* 910 cells were transformed to ampicillin resistance with 15 µl aliquots (out of 50µℓ totals) of the chimaeric DNA molecules recovered from the gel, and with 20 µℓ aliquots (out of 100µℓ totals) of the ligation products derived from the phosphatase-treated pBR322. Ampicillin-resistant (Ap^R) transformants were transferred, using sterile toothpicks, in a regular grid pattern ("patched") onto duplicate L Ap and L Ap+Tc plates. Tc^s transformants which failed to grow on the L Ap Tc plate were readily identifiable on the duplicate L Ap "master" plate.

2.2.11. ^{32}P-labelling of DNA by Nick-translation.

In vitro labelling of DNA was based on the method of Rigby *et al.* (1977) and was performed using the Amersham nick-translation kit as described in its instructions. A typical reaction for labelling DNA to a specific activity of 10^8dpm/µg contained 10µℓ (∿0.5µg) of DNA, 10µℓ of solution 1 (100 µM dNTP;5x nick-translation buffer), 5µℓ(50µCi; 125pmol) of α-^{32}P-dCTP, 5µℓ of solution 2 (2.5u DNA polymerase 1,50 pg DNase 1)and 20µℓ of water. The mixture was incubated at 15°C for 2hr after which SDS was added from a 10% (w/v) stock solution to a final concentration of 0.1%. The labelled DNA was separated from the unincorporated label by chromatography on a column of Sephadex G50 (superfine grade) equilibrated and eluted with 150 mM NaCl, 10mM EDTA, 50mM Tris-HCl pH 7.5, 0.1% SDS. A 1µℓ aliquot of the collected DNA eluate ∿1.6ml) was dispersed in 50ml of scintillation fluid (3.37g PPO/667 ml toluene, 333 ml Triton X-100 per litre), and the radioactivity was determined using a Packard (PL Tri-carb Prias) liquid scintillation counter.

2.2.12. ^{32}P-5' end labelling of RNA.

Poly(A)$^+$ molecules were labelled by the method of Bedbrook *et al.*(1980). 5.0µg of RNA were subjected to partial hydrolysis by heating at 95°C for 5 min in 10µℓ of 5mM Tris-HCl pH 9.5, 10mM EDTA, 0.1mM spermidine, and then cooled on ice. 5µℓ of 4X kinase buffer (200 mM Tris-HCl pH 9.5, 50mM $MgCl_2$, 40mM DTT, 20% glycerol), 5µℓ (50µCi) of γ-^{32}P-ATP, and 1µℓ(5u) of T4 polynucleotide kinase were added and incubated at 37°C for 30min. 1µℓ of 10 mM ATP was added and the incubation was continued at 37°C for 30min. The mixture was diluted to 200 µℓ with TE buffer, and after the addition of 40µg of *E.coli* tRNA,it was extracted with phenol, ethanol precipitated and redissolved in TE buffer. The labelled RNA was purified by electrophoresis through a column of a 1% agarose gel in a 1 ml pipette tip essentially as described by Grunstein and Wallis (1979).

2.2.13. Processing of Bacteria for *in situ* Colony Hybridisation.

The procedure used was based on the method of Grunstein and Wallis (1979) and included modifications described by Maniatis *et al.*(1982). Bacterial colonies were "patched" in replicate onto a "master" agar plate containing selective antibiotics, and onto nitrocellulose filter discs (82mm) overlaid on selective agar plates. The colonies were grown to ∿1mm diameter at 37°C (or 30°C with clones harbouring temperature-inducible expression plasmids), at which stage the "master" plate was stored at 4°C. With colonies harbouring amplifiable plasmids, filters were sometimes transferred to plates containing chloramphenicol (170µg/ml),and incubated overnight at 37°C (or 30°C). The colonies were processed for hybridisation by sequentially placing the filter, colony side up, for 5min. periods on stacks of 4 sheets of 3MM paper saturated with the following solutions : i) 10% SDS; ii) 0.5M NaOH, 1.5M NaCl; iii) 0.75M Tris-HCl pH 7.5, 1.5M NaCl; iv)3 x SSC (0.45M NaCl, 0.045M Na Citrate pH 7.0). The filter was air-dried, then baked between 2 sheets of 3MM paper for 2hr at 80°C under vacuum. Screening of the clones with a ^{32}P-labelled probe was as described in section 2.2.15. Positive clones were identified by autoradiography, and the relevant colonies were selected from the "master" plate.

2.2.14. Southern Blotting : Transfer of DNA from Agarose Gels to Nitrocellulose Paper.

The procedure used was modified from the method originally described by Southern (1975). DNA fragments, fractionated by gel electrophoresis were denatured by agitating the gel in denaturing buffer (1.5M NaCl, 0.5M NaOH, 1.0mM EDTA) for 30min with one change of buffer. The gel was then neutralised by shaking for 30 min in neutralising buffer (3.0M NaCl, 0.5M Tris-HCl pH 7.0, 1mM EDTA) with one change of buffer, and then equilibrated in 20x SSC (3.0M NaCl, 0.3M Sodium citrate adjusted to pH 7.0 with HCl) for 15 min. Capillary blotting of the DNA was performed by overlaying the gel with a nitrocellulose filter and absorbent towels as described by Maniatis *et al.* (1982) with the following modifications : i) 20x SSC was used as the transfer buffer; ii) the nitrocellulose filter was prewetted by floating on the surface of distilled water and was then submerged in 20x SSC for 15 min. ; and iii) the nitrocellulose filter was overlaid with a sheet of 3MM paper wetted in 20x SSC, three sheets of dry 3MM paper and three layers (∿3cm) of disposable baby nappies (Boots, Nottingham, U.K.). The transfer was allowed to proceed for at least 15hr at 4°C.

2.2.15. Hybridisation of ^{32}P-labelled Probes to Filter-bound DNA.

This technique was used to detect DNA which had been transferred to nitrocellulose filters by *in situ* lysis of bacterial colonies (section 2.2.13) or by Southern blotting (section 2.2.14). All filter washes and the hybridisation itself were carried out in heat-sealed plastic bags submerged in a shaking waterbath at 50-68°C depending on the desired stringency. The filter was equilibrated first in 3X SSC (1-2ml per cm^2 of filter) for 15min., then in 3X SSC, 10X Denhardt's solution (o.2%(w/v) each of BSA, polyvinylpyrrolidone, and Ficoll 400) for an additional 15 min. It was then prehybridised in 3X SSC, 10X Denhardt's solution containing 100 µg/ml each of denatured herring sperm DNA, ATP and poly(dA)(0.1-0.5 ml per cm^2 of filter) for 1-2hr. DNA probes were denatured by boiling for 10 min. before addition to the prehybridisation solution whereas RNA probes did not require heat-denaturation before use. Hybridisation was usually allowed to proceed overnight though hybridisation times as short as 5hr were successfully employed on occasion. After hybridisation, the hybridisation mixture was poured off and stored at -20°C for re-use later. If low hybridisation stringency was desired, the filter was washed for four 15min. periods in 3X SSC buffer at 50°C. For higher stringencies, the ionic strength of the wash solutions was progressively decreased. For very high stringency, for example, the filter was washed at 68°C for two 15min. periods in 3X SSC, two 15 min. periods in 1X SSC and finally two 30 min. periods in 0.1X SSC. It was then blotted dry on a sheet of 3MM paper and autoradiographed.

2.2.16. Autoradiography.

Autoradiography was used to locate ^{32}P-labelled nucleic acids on nitrocellulose filters and in polyacrylamide gels. The following manipulations were done in a dark-room under a red safelight. A sheet of preflashed X-ray film (Fuji RX; Laskey and Mills, 1975) and an intensifying screen (Dupont Cronex Lighting-Plus) was placed over the sample, sandwiched between two glass plates, and secured with rubber bands. The assembly was placed in three, black plastic bags to exclude light, and exposure (-80°C or room temperature) was varied from 30min. to several weeks depending on the sample. The film was developed in Kodak X-Omat developer at room temperature for 3-8min., washed for 1 min. in water, fixed in Kodak fixer for 3min., washed again in water for 30min. and air-dried at room temperature.

2.2.17. Fluorography.

SDS-polyacrylamide gels containing fractionated, tritium-labelled polypeptides were processed for fluorography by a method adapted from Bonner and Laskey (1974). The gels were soaked for 30 min. with constant agitation in DMSO, followed by a second 30min. immersion in fresh DMSO. They were then soaked for 3hr in a solution of PPO (30% (w/v) in DMSO), and then for 1hr in 30% (v/v) methanol. The gels were dried under vacuum (Bio. Rad Model 224 gel slab dryer) between two layers of cellophane (W.E. Cannings, Avonmouth, Bristol, U.K.) and exposed at -80°C to presensitised Fuji RX film as described in the preceding section.

2.2.18. Restriction Mapping of Cloned cDNAs.

Samples of plasmid DNA (∿0.5μg) were digested initially with enzymes that recognised hexanucleotide sequences and thus likely to cleave the insert infrequently. The restricted DNA was electrophoresed on a 0.5-1% agarose gel and the positions of cleavage sites in the inserts were deduced from the sizes of the restriction fragments and a knowledge of the target sites in the pBR322 vector (Sutcliffe, 1978). For more detailed mapping, digestions were carried out simultaneously with enzymes recognizing tetranucleotide sequences and with BamHI. An inspection of the restriction fragments on an agarose gel immediately indicated whether the BamHI-excisable, cDNA insert had been cleaved. To map the sites of enzymes which cleaved the insert, additional multiple digestions were performed and analysed by agarose gel electrophoresis until all the cleavage sites were unambiguously assigned to internally consistent positions.

2.2.19. Characterisation of Cloned cDNAs by Hybrid-selected Translation.

Activated diazobenzoyloxymethyl (DBM) paper was prepared by the method of Alwine *et al.*(1977). Recombinant plasmids were restricted with BamHI to excise their cDNA inserts, and 25μg aliquots were denatured and bound to discs of DBM paper essentially as described by Smith *et al.*(1979). Poly(A)$^+$ RNA (50μg) prepared from 14-day-old pea cotyledons was hybridised to the immobilised plasmids for 3hr at 37°C in 550 μℓ of hybridisation buffer (20mM PIPES pH 6.4, 0.9M NaCl, 0.2% SDS, 1mM EDTA, 50% formamide, 300 units/ml RNasin). The filters were then washed twice with 500μℓ aliquots of hybridisation buffer and once with 500μℓ of 20mM NaCl, 8mM sodium citrate, 1mM EDTA,0.2% SDS, 50% formamide, 300 units/ml RNasin, at 37°C for 30min. periods. The specifically bound RNA was eluted by incubating at 37°C for 30min. in 100μℓ of 20mM PIPES pH 6.4, 1mM EDTA, 0.5% SDS, 90% formamide, 300 units/ml RNasin. The eluted RNA was recovered by ethanol precipitation and translated in the rabbit reticulocyte lysate system using tritium-labelled leucine (0.5μCi/ml) as the radioactive label (Croy *et al.* 1980a). After a 1hr incubation at 37°C, translation products were analysed by SDS-PAGE on 17% gels followed by fluorography as previously described.

2.2.20. DNA Sequencing.

The dideoxynucleotide-terminated, nick-translation method originally described by Maat and Smith (1978) and later modified by Seif *et al.* (1980) was used for sequencing. DNA fragments with single labelled 5' termini were generated (Maxam and Gilbert, 1980) for use in the nick-translation reactions. The steps in the sequencing protocol are outlined below.

2.2.20.1. Preparation of DNA Fragments with Single Labelled 5' Termini.

12-15μg of restriction-mapped DNA (usually recombinant pBR322 plasmids) were digested to completion with a restriction enzyme(s) which gave 5' protruding termini. Where difficulty was experienced with end-labelling certain DNA preparations, the required fragment(s) was purified on an agarose gel before proceeding with the next reaction. The restricted DNA fragments were 5'-dephosphyorylated by treatment with alkaline phosphatase (section 2.2.4.2.), and end-labelled by treatment with 20 units of T4 polynucleotide kinase at 37°C for 45 min in 25μℓ of 50mM Tris-HCl pH 7.6, 10mM $MgCl_2$, 5mM DTT, 0.1mM EDTA, 1.5mM spermidine, 0.8μM ATP and 1μM (125μCi) γ-^{32}P-ATP. The reaction was terminated by the addition of 200μℓ of a 2.5M ammonium acetate solution. 1μℓ of a 1μg/μℓ solution of *E.coli* tRNA was added and the DNA was ethanol precipitated. Most of the unincorporated label was removed by two successive resuspensions of the DNA in 300μℓ aliquots of 0.3M sodium acetate followed by ethanol precipitations. To generate fragments labelled at only one 5' terminus, the DNA was digested with a second restriction enzyme(s) that cut the fragment(s) asymmetrically. A small sample (∿6%) of the undigested DNA was kept for use as a size marker and control on the preparative gel. The restricted DNA was ethanol precipitated and redissolved in a small volume (∿10μℓ) of TE buffer. Both the restricted sample and the control sample were fractionated on a 5% polyacrylamide gel (section 2.2.7.2.), and the labelled fragments were visualised by autoradiography at room temperature for 20-40 min. Using the autoradiogram as a guide, small sections containing the required fragments were excised from the gel. The fragments were recovered by the crush-soak procedure described by Maxam and Gilbert (1980) except that the eluate was collected by being 'blown out' with a stream of N_2 gas instead of by centrifugation, and the crushed gel was washed with 150μℓ aliquots of the elution buffer (generally 2 to 3 washes) until a Geiger counter indicated that most of the labelled DNA had been recovered. The labelled DNA was ethanol-precipitated, resuspended in 400 μℓ of TE buffer, reprecipitated with ethanol, and finally redissolved in 20μℓ of water.

2.2.20.2. Dideoxynucleotide-terminated Nick-translation Sequencing Reactions.

Solutions of the following NTP mixtures were used in the forwards (F) and backwards (B) sequencing reactions (see Seif *et al.*,1980).

FG	1mM ddGTP, 200μM dATP, 200μM dTTP, 200μM dCTP
FA	1mM ddATP, 200μM dGTP, 200μM dTTP, 200μM dCTP
FT	1mM ddTTP, 200μM dGTP, 200μM dATP, 200μM dCTP
FC	1mM ddCTP, 200μM dGTP, 200μM dATP, 200μM dTTP
BG	1mM ddGTP
BA	1mM ddATP
BT	1mM ddTTP

Narrow plastic tubes (narrow enough to fit into small scintillation vials for Cerenkov counting) were labelled FG,FA, FT, FC, BG, BA, BT and BC. 1.1μℓ aliquots of the appropriate NTP mixture were dispensed into each tube which was immediately placed on ice. To the 20μℓ of the labelled DNA were added 10μℓ of 5X Seif buffer (33mM Tris-HCl pH7.5, 33mM $MgCl_2$, 10mM DTT, 10mM NaCl), 8μℓ of endonuclease-free DNA Polymerase 1 (5 units/μℓ), and 2μℓ of DNase 1 (100ng/ml). 4.5 μℓ of this DNA-enzyme mixture were added to each of the NTP mixtures. The tubes were spun briefly in a microcentrifuge and were incubated at 37°C for 30min. 1μℓ of 0.1M EDTA was added to each of the B tubes and the tube contents were transferred to the corresponding F tube. Small holes were made in the caps of the F tubes which were then stored at -80°C for 30min. The samples were then lyophilised for at least 1hr.

2.2.20.3. Electrophoresis of Sequencing Samples.

10μℓ of formamide dye solution (10mM NaOH, 1mm EDTA, 80%(v/v) formamide, 1mg/ml bromophenol blue, 1mg/ml xylene cyanol) were added to each sample which was then heated at 90°C for 5 min.

The radioactivity of the samples was determined in a scintillation counter (Packard Tricarb Prias PL). The tubes were subsequently stored on ice. Just prior to loading, each sample was heated at 90°C for 1 min., then returned immediately into ice. 2.5µℓ aliquots of each sample, were electrophoresed per track on thin 6 or 8% polyacrylamide-urea gels (see section 2.2.7.3).

When electrophoresis was completed, the gel was dried at 80°C for 1hr under vacuum (Bio-Rad Model SE 1125B gel slab dryer) and autoradiographed at -80°C. An estimate of the required exposure was obtained from the formula :

$$\text{exposure (hr)} = \frac{1.3 \times 10^6}{\text{radioactivity of sample (cpm) per gel track}}$$

Usually, an exposure of 24-60hr was required. Highly radioactive samples were often re-autoradiographed without an intensifying screen as this procedure gave sharger bands on the autoradiograph. The required exposure was then 10-15 times as long as with a screen.

2.2.21. Construction of Vicilin Expression Plasmids.

The construction of vicilin expression plasmids involved essentially the subcloning of vicilin cDNA fragments into various expression vectors. In some constructions, the vector DNA (1.5-3µg) was cleaved with a suitable enzyme, 5'-dephosphorylated with alkaline phosphatase, and then ligated in 3 to 5-fold molar excess to a cDNA fragment with compatible ends. In other cases, the cDNA fragment was 5'-dephosphorylated and ligated in 10 to 20-fold molar excess to the vector (0.2-0.8 µg) restricted with an appropriate enzyme. Both these strategies were designed to minimize the proportion of non-recombinant, recircularised plasmid molecules formed in the ligation reaction. The choice between them was largely dictated by the availability of the participating DNA species and details of the actual methods used are given in the "Results".

2.2.22. Analysis of *E.coli* Expression Systems by SDS-PAGE.

Cultures of cells harbouring λP_L-expression plasmids were grown at 30°C in L broth containing 50μg/ml ampicillin. Aliquots of the culture were withdrawn at various cell densities (see section 3.5.2.) and were induced by incubation at 42°C using two slightly different procedures (section 3.5.2.). Incubation of the remainders of the cultures was continued at 30°C. At the end of the induction, 1.5 ml aliquots of both induced (42°C) and uninduced (30°C) cultures were pelleted at 12000g for 15min., resuspended in 100μℓ aliquots of sample buffer (20mM Tris-HCl ph 6.8, 2%(w/v) SDS, 2%(v/v) mercaptoethanol, 10%(v/v) glycerol), and heated in a boiling water bath for 5 min. Insoluble cell debris was spun down at 12000g for 10min and 30μℓ aliquots of the supernatants were subjected to SDS-PAGE. The proteins were visualised by staining with Kenacid blue and by immunological screening of Western blots.

2.2.23. Western Blotting : Electrophoretic Transfer of Proteins from SDS-polyacrylamide Gels to Nitrocellulose Paper.

Proteins were transferred from SDS-polyacrylamide gels to nitrocellulose paper in a Trans-Blot Cell (Bio. Rad Laboratories) containing deaerated transfer buffer (25mM Tris base, 192mM glycine, 20% (v/v) methanol, pH 8.3) at 7V/cm for at least 12hr (Towbin *et al.* 1979; Burnette, 1981).

2.2.24. Processing of Bacteria Harbouring Expression Plasmids for *in situ* Colony Immunoassay.

The procedure of Helfman *et al.*(1983) was used. Bacterial colonies were "patched out" and grown at 30°C on a selective, "master" agar plate and in duplicate on a nitrocellulose filter overlaid on a selective plate. When the colonies had grown to 1-2mm in diameter, the "master" plate was stored at 4°C while the colonies on the filter were incubated at 42°C for 2.5hr. The cells were lysed *in situ* by suspending the filter in an atmosphere saturated with chloroform vapour for 20min. It was then transferred to a petri dish containing 10ml of 50mM Tris-HCl pH 7.5, 150mM NaCl, 5mM $MgCl_2$, 30mg/ml BSA, 1μg/ml DNase 1 and 40μg/ml lysozyme, and was incubated overnight at room temperature with gentle agitation.

2.2.25. Immunological Detection of Filter-bound Vicilin Polypeptides.

This technique, based on the method of Towbin *et al.*(1979) was used to screen proteins which had been transferred to nitrocellulose filters by Western blotting (section 2.2.23.) or by *in situ* lysis of bacterial colonies (section 2.2.24). The Western-blotted filter was incubated with gentle agitation in 100ml of Tris-saline buffer (20mM Tris-HCl pH 7.4, 0.9% (w/v) NaCl) containing 5% (w/v) BSA for 1hr at 40°C in a heat-sealed plastic bag. The following steps apply to both Western-blotted and colony screen filters. The filter was rinsed briefly at 25°C in 100ml of Tris-saline buffer in a plastic bag. It was then incubated for 2-3hr with 30μℓ of affinity-purified, rabbit antivicilin IgG diluted into 10ml of Tris-saline buffer, 5% BSA in a petri dish (at room temperature) for bacterial colony screens, or 40μℓ of antibody diluted into 30ml of buffered BSA in a plastic bag (at 25°C) for Western blots. The filter was washed for ∿45min in 4 changes of Tris-saline buffer (100-200ml per wash) at 25°C. It was then incubated for 1hr at room temperature with 30μℓ of swine, peroxidase-conjugated, antirabbit IgG (Orion Diagnostica, Helsinki, Finland) diluted as above for colony screens, or with 40μℓ of antibody at 25°C for Western blots. Filters were then washed in Tris-saline buffer as above. 20ml of a solution of 4-chloro-1-napthol (3mg/ml in methanol) were diluted into 100ml of Tris-saline buffer, and 40μℓ of 30%(v/v) hydrogen peroxide solution were added. The nitrocellulose filters were placed in this mixture and kept in the dark. The reaction was stopped after suitably intense staining had been achieved (20-30 min) by washing the filter thoroughly in water. The blots were dried between sheets of 3MM paper and stored in the dark.

3. RESULTS.

3.1. Construction of cDNA Library.

3.1.1. Synthesis and Size Fractionation of ds-cDNA.

Double-stranded cDNA (ds-cDNA) was synthesised from pea cotyledon poly(A)$^{+}_{+}$ mRNA as previously described (section 2.2.10.2.) The mRNA used for the cDNA synthesis had previously been characterised and shown to be enriched in species encoding the major seed storage proteins (Evans *et al.*,1980). The ds-cDNA molecules obtained were predominently in the size range of 0.2 - 3.0 Kb as estimated by agarose gel electrophoresis (Fig.1). To maximize the cloning of long cDNA molecules, two fractions comprising cDNAs of lengths ∿1-2 Kb and >2Kb were recovered from the gel. It was calculated from ^{3}H-incorporation that ∿5μg of the former and ∿2μg of the latter size fraction were recovered.

3.1.2. Trial Phosphorylation, Ligation and Restriction of BamHI Linkers.

Decameric, BamHI linkers (Collaborative Research) with 5'-hydroxyl ends had to be phosphorylated prior to ligation. The ability of the linkers to undergo phosphorylation, ligation and subsequent restriction with BamHI was tested in a pilot experiment prior to being used for cDNA cloning. The linkers were phosphorylated with γ-^{32}P- ATP, self-ligated and then digested with BamHI (Section 2.2.10.3.). Samples were analysed by PAGE followed by autoradiography, and the results are shown in Fig.2. The tracks containing the self-ligated linkers show that the phosphorylation and blunt-end ligation worked efficiently since labelled, oligomeric linker molecules (up to 10-mers) are visible on the autoradiograph. In the initial trial restriction, a sample of the ligated linkers was digested with 5.4 units of BamHI which gave only partial restriction (Fig.2, Track A2). The digestion was repeated on a second aliquot of the ligated linkers using 11 units of BamHI which then gave complete restriction (Fig.2, track B2). Identical digestion conditions were subsequently used for the digestion of linkered molecules in the cDNA cloning experiment.

3.1.3. Ligation of cDNA to pBR322 Vector.

BamHI-digested, linkered cDNA molecules from both size fractions were ligated in separate reactions to phosphatase-treated and

FIGURE 1. Size fractionation of double-stranded cDNA from pea cotyledon mRNA, by electrophoresis through a 0.5% agarose gel.

Tracks:

1) pBR322 restricted with AluI;

2) and 3) pea ds-cDNA;

4) pBR322 restricted with HindII.

FIGURE 2. Autoradiographs of ^{32}P-phosphorylated, ligated and restricted BamHI linkers separated by PAGE.

Tracks:

A1) and B1) ligated linkers;

A2) ligated linkers restricted with 5.4 units of BamHI;

B2) ligated linkers restricted with 11 units of BamHI.

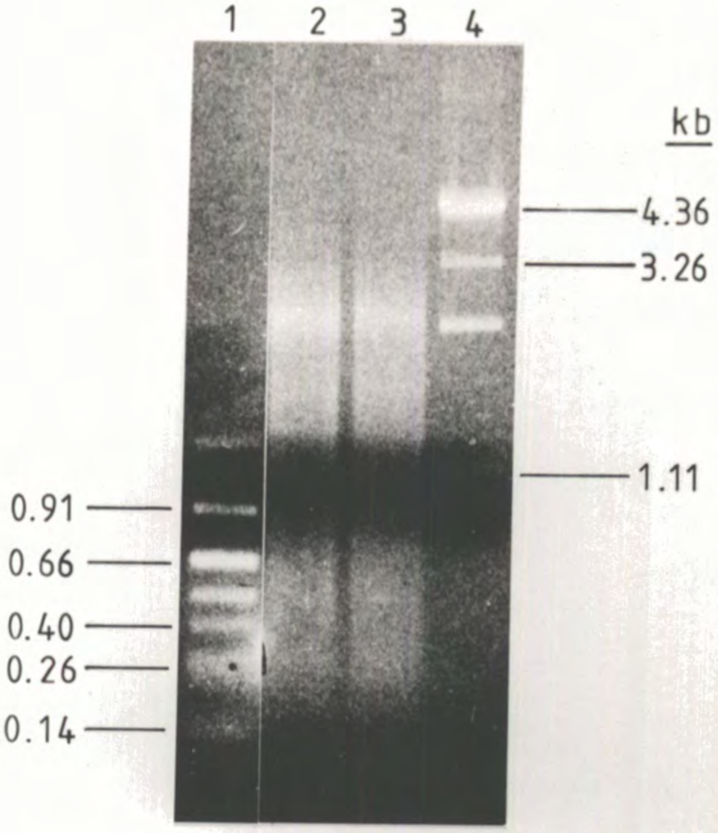

1
2
3
4
kb
4.36
3.26
1.11
0.91
0.66
0.40
0.26
0.14

Fig. 1

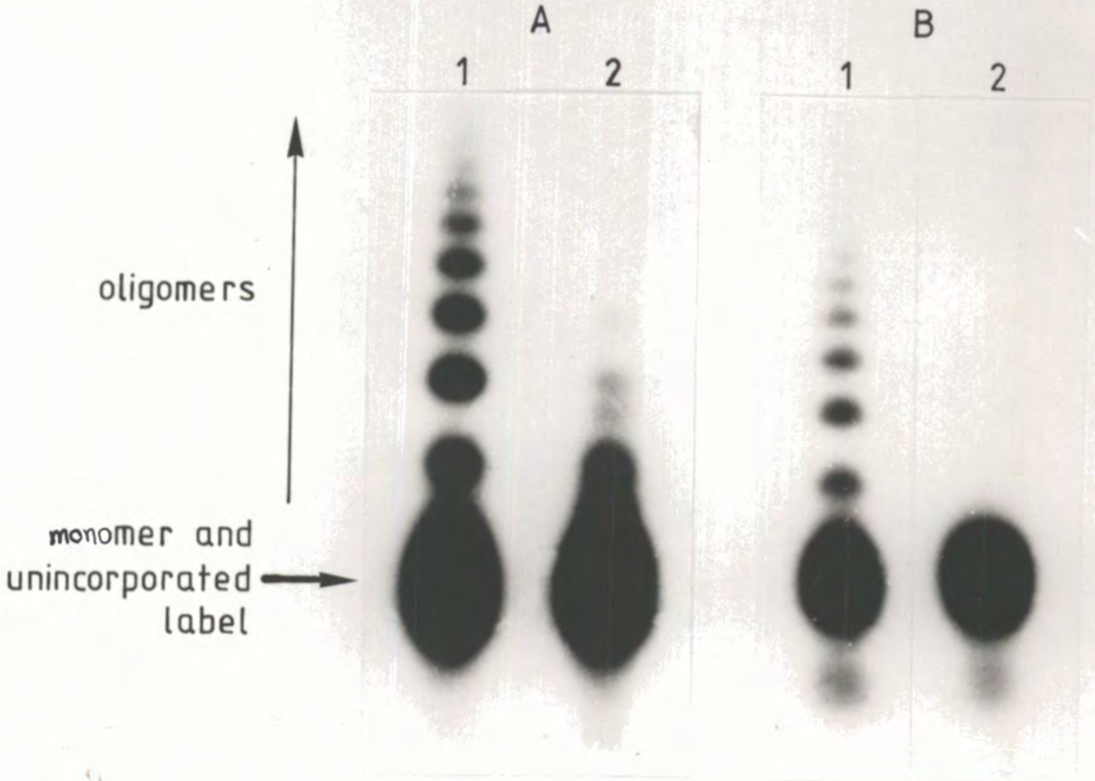

A
1
2
B
1
2
oligomers
monomer and unincorporated label

Fig. 2

non-phosphatase-treated, BamHI-linearised pBR322. (section 2.2.10.6). Agarose gel electrophoresis of a small aliquot of the ligation products of the cDNA and phosphatased pBR322 showed that no significant amounts of oligomeric pBR322 molecules were present (gel not photographed). By contrast, oligomeric forms of the vector were prominent among the gel-fractionated products of the ligation between the cDNA and the nonphosphatased plasmid (Fig.3). Vector-cDNA hybrids presumably made up the faint smears visible between the monomeric and dimeric vector bands. These hybrid molecules were recovered from the gel for transformation of an *E.coli* host.

3.1.4. Screening of Bacterial Transformants for Recombinant Plasmids.

BamHI-restricted, linkered cDNAs were ligated into the BamHI site in pBR322 and transformed into *E.coli* 910. Resistance to ampicillin (Ap^R) allowed identification of transformants, and sensitivity to tetracycline (Tc^S) caused by insertional inactivation of the Tc^R gene in pBR322 identified clones harbouring recombinant plasmids. The numbers of Ap^R colonies obtained after transformation with various vector-cDNA samples, and the numbers of Tc^S colonies subsequently identified are shown in Table 6. The numbers of Ap^R transformants obtained from the phosphatased plasmids were 5-6-fold higher than the numbers obtained from the non-phosphatased plasmids. The proportion of Tc^S transformants was very low irrespective of the origin of the DNA used for transformation —— out of 3530 Ap^R transformants screened, only 129 (3.7%) were found to be Tc^S.

Table 6. Results of Transformation of *E.coli* cells with pBR322-cDNA Ligation Products.

Transforming DNA$^{(a)}$	No. of .Ap^R transformants obtained.	No. of Ap^R colonies screened for Tc-sensitivity.	No. of Tc^S colonies obtained.
A	∿3500	930	52
B	∿4500	1250	49
C	∿ 800	750	23
D	∿ 700	600	5

Key :Ap^R = ampicillin-resistant; Tc^S = tetracycline-sensitive.
a. A=ligation products of 1-2Kb cDNA and phosphatased vector ;
B=ligation products of >2Kb cDNA and phosphatased vector;
C=hybrid 1-2Kb cDNA-vector molecules recovered from gel;
D=hybrid > 2Kb cDNA-vector molecules recovered from gel.

FIGURE 3. Ligation products of cDNAs and BamHI - linearised pBR322, fractionated on a 0.5% agarose gel.

Tracks:

1) λNM258 DNA restricted with EcoRI;

2) 1-2 Kb cDNA fraction ligated to vector;

3) $\geqslant$2 Kb cDNA fraction ligated to vector.

FIGURE 4. Autoradiographs of nitrocellulose filters bearing *E.coli* colonies hybridised to ^{32}P-labelled mRNA and cDNA probes as follows:

A) pea cotyledon poly(A)$^{+}$ mRNA;

B) pDUB3 legumin cDNA insert;

C) pDUB4 vicilin cDNA insert;

D) pAD2-1 vicilin cDNA insert.

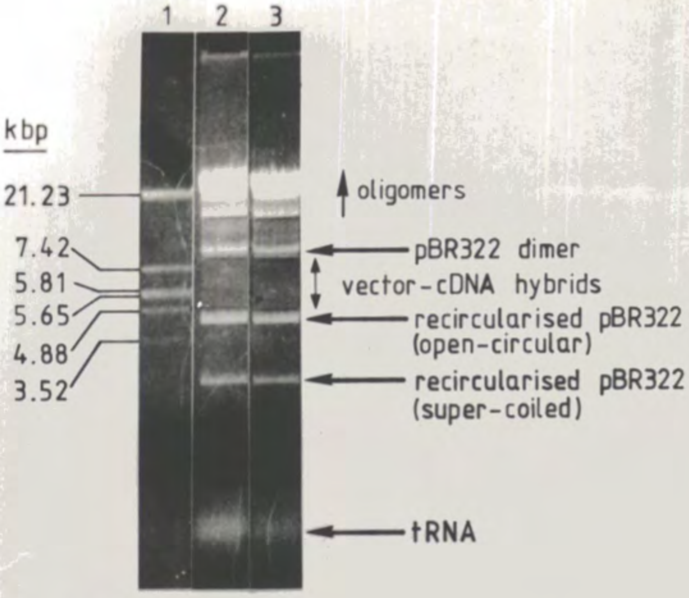

1
2
3
kbp
21.23
7.42
5.81
5.65
4.88
3.52
oligomers
pBR322 dimer
vector-cDNA hybrids
recircularised pBR322
(open-circular)
recircularised pBR322
(super-coiled)
tRNA

Fig. 3

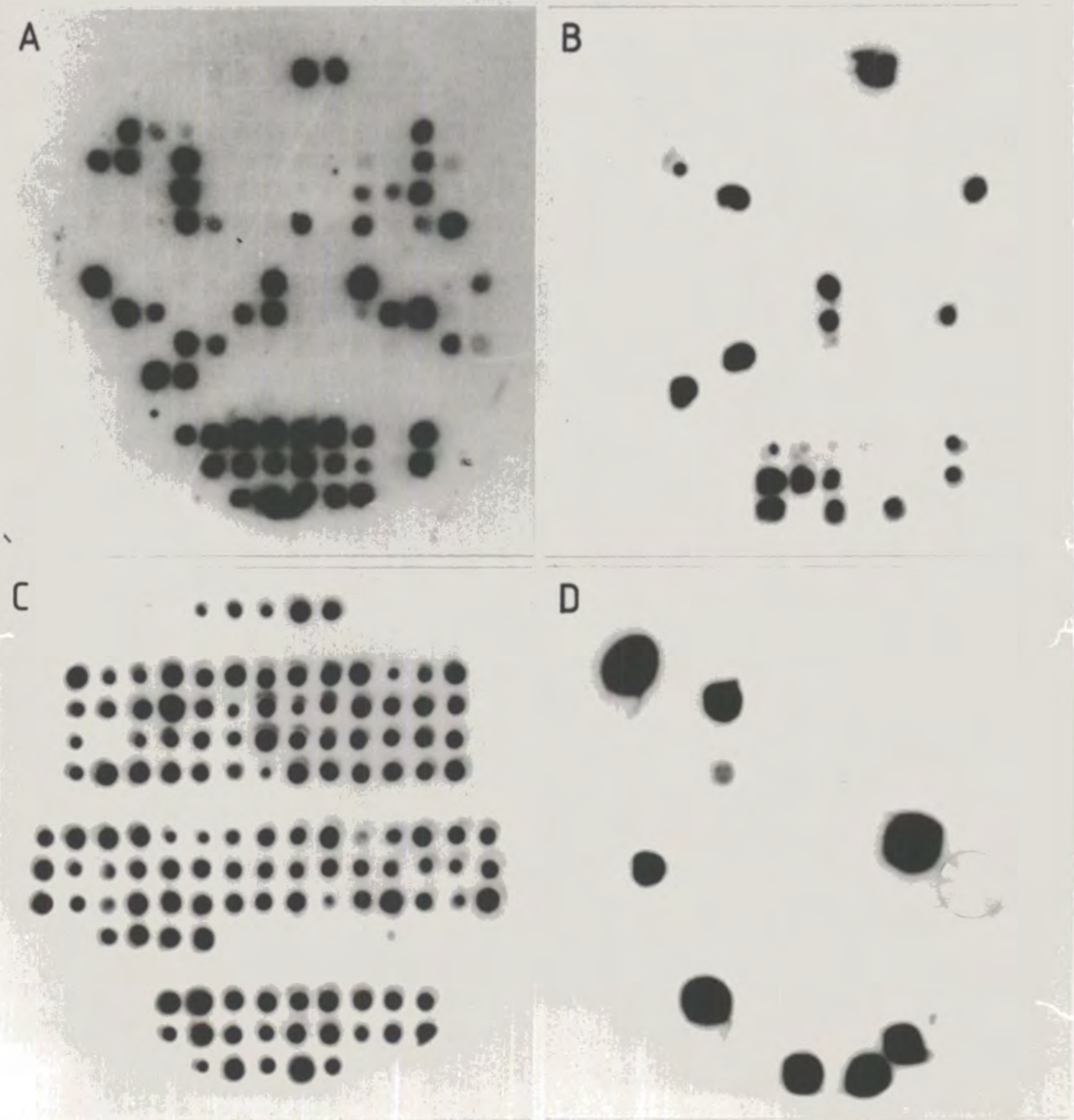

A
B
C
D

Fig. 4

3.2. Characterisation of cDNAs from the Clone Library.

3.2.1. Notes on Previously Characterised cDNA Clones Used for Screening of the Library.

Previously isolated and characterised legumin and vicilin cDNA clones were used to screen the present cDNA library. The legumin clone pDUB3, was originally designated pRC2.11.7 (Croy *et al.*,1982) before renaming systematically (*D*urham *U*niversity *B*otany). The 830bp cDNA insert encodes the basic subunit of a legumin molecule and thirty C-terminal residues of the acidic subunit.

The vicilin clones pDUB2 and 4 (Lycett *et al.*,1983a) were originally designated pRC2.2.1 and 2.2.10 respectively (Croy *et al.*, 1982). The 910bp insert from pDUB2 encodes part of a 50000-*M*r vicilin subunit, whereas the 210bp insert from pDUB4 encodes part of a 47000-*M*r subunit.

3.2.2. Identification of cDNA Clones by Colony Hybridisation.

In order to classify the cloned cDNAs into storage protein-specific groups, the Ap^R Tc^S transformants were grown on replicate nitrocellulose filters and screened by *in situ* colony hybridisation with ^{32}P-labelled RNA and cDNA probes. The results are summarised in Table 7. 59 of the 129 clones screened hybridised to a poly(A)$^+$ RNA probe under high stringency conditions (0.1 x SSC at 65°C; Fig. 4A). Of these, 23 hybridised under similar high stringency criteria to the legumin cDNA excised from pDUB3 (Croy *et al.*,1982; Fig.4B).

One of the pDUB3-hybridising cDNAs, pAD4-4, was shown by subsequent restriction mapping (Fig. 6) to extend by ∿300bp beyond the 5' end of pDUB3. The 360bp (BamHI-BglI) 5' terminal fragment of pAD4-4 (see Fig.6) was used to rescreen the cDNA library. No additional positives were scored, and the relative intensities of hybridisation were generally less than that seen with the pDUB3 probe (see Table 7). This indicated that most of the cloned legumin cDNAs did not extend as far as the pAD4-4 insert towards the 5' end of the legumin mRNA. Notable exceptions were pAD10-3, 10-4, 10-5 and 10-6 which hybridised more strongly to the pAD4-4 5' terminus than to pDUB3.

Table 7. Characterisation of cDNA Clones [a] by Colony Hybridisation and Sizes of Inserts.

			Hybridisation[b] to labelled probes			
Clone [c,d]	pDUB No.[c]	Approx.size of insert (bp)	mRNA[e]	pDUB3[f] insert	pAD4-4 350bp 5' region	pAD2-1 insert
pAD1-4		830	+++	+++	+	
pAD1-5		830	+++	+++	+	
pAD2-1	pDUB9	1430	+++			+++
pAD2-2		ND	+			
pAD2-3		ND	+			
pAD2-11		1020	+++			
pAD3-1		ND	+++			
pAD3-2		1650[g]	+++	++	+++	
pAD3-4	pDUB7	1080	+++			++
pAD3-10		ND	+			
pAD3-12		820	++			
pAD3-13		ND	+			
pAD4-4	pDUB6	1120	+++	+++	+++	
pAD4-10		?[h]	+			
pAD4-11		ND	+			
pAD4-12		1000+530[g]	+++	+++		
pAD5-4	pDUB28	1200	+++			+
pAD5-5	pDUB29	530	+			+
pAD5-8		?[h]	++			+
pAD5-10		980	++			
pAD5-12		ND	+			
pAD5-13		800	+++			
pAD6-2		1850	+++			
pAD6-8		1000+530[g]	+++	+++		
pAD6-11	pDUB10	1950	+++			+++
pAD6-15		<300	+	+		
pAD7-3	pDUB31	860	+++			++
pAD7-4		560	++			+
pAD7-7		560	++			+
pAD7-8		830	++	+++	+	
pAD7-11		ND	+			
pAD7-12		830	+++	+++	+	
pAD7-13	pDUB11	1820	+++			+++
pAD8-5		1000+530[g]	+++	+++		

Table 7 continued...

Clone	pDUB No.	Approx.size of insert (bp)	mRNA	pDUB3 insert	pAD4-4 350bp 5' region.	pAD2-1 insert.
pAD8-6		560	++			+
pAD8-14	pDUB30	560	++			+
pAD8-15		ND	+			
pAD9-2		1850	+++			
pAD9-3		830	+++	+++	+	
pAD10-1		$?^{(h)}$	++			++
pAD10-2		980	+++			
pAD10-3		950	+++	+	+++	
pAD10-4		950	+++	+	+++	
pAD10-5	pDUB8	950	+++	+	+++	
pAD10-6		950	+++	+	+++	
pAD10-7		1000	++			
pAD10-9		830	+++	++	+	
pAD11-2		1000	++			
pAD11-3		830	+++	+++	+	
pAD11-4		830	+++	+++	+	
pAD11-5		830	+++	++	+	
pAD11-6		ND	++			
pAD11-7		$?^{(h)}$	+			++
pAD11-9		830	+++	++	+	
pAD12-1		830	++	+++	+	
pAD12-2		$?^{(h)}$	++			++
pAD12-3		830	+++	+++	+	
pAD12-4		$?^{(h)}$	++			++
pAD12-5		830	+++	+++	+	

a. Only clones which hybridised to the mRNA probe are included in the table.

b. Approximate, relative intensities of hybridisation are indicated on a scale of "+" to "+++"

c. Notes on nomenclature : all the recombinant plasmids from the cDNA bank were initially identified by a pAD number according to their original positions on the colony hybridisation filters. This system of nomenclature is used throughout the thesis to differentiate clearly between these clones and plasmids obtained from other sources. Some of the pAD plasmids were subsequently given pDUB (*D*urham *U*niversity *B*otany) numbers as indicated, and will be referred to by these pDUB designations in papers for publication.

d. Clones pAD1-4 and pAD1-5 were obtained from transformation with DNA "D"; pAD2-1 to pAD5-13 from DNA "A"; pAD6-2 to pAD9-3 from DNA "B"; and pAD10-1 to pAD12-5 from DNA "C" (see footnotes to Table 6 for details of A,B,C and D).

e. mRNA = poly(A)$^+$ RNA prepared from pea cotyledons 14 days after flowering.
f. Formerly designated pRC2.11.7 (Croy *et al.* 1982).
g. See section 3.2.4.
h. This plasmid contained only one BamHI site.

Screening of the library with the vicilin cDNAs from pDUB2 and 4 (Lycett *et al.*,1983a) gave very high backgrounds on the autoradiographs making it impossible to identify clones which specifically hybridised to these probes (Fig.4C). To initially identify vicilin cDNAs, plasmid mini-preps from a selection of clones were digested with BamHI, fractionated by agarose gel electrophoresis, blotted onto nitrocellulose filters and probed with labelled pDUB2 and 4 inserts (see section 3.2.3). One of the vicilin cDNAs, pAD2-1, identified by Southern blot hybridisation and subsequently fully characterised, was used to rescreen the clone bank by colony hybridisation. A satisfactory autoradiograph with a "clean" background was obtained under moderate stringency conditions (Fig. 4D). 16 clones hybridised to the pAD2-1 excised insert, of which 14 had previously been identified as vicilin clones by their Southern blot hybridisation to the pDUB2 and 4 inserts (Lycett *et al.*,1983a; section 3.2.3.).

The library was also screened with a labelled cotyledon rRNA probe under high stringency conditions. None of the clones hybridised to that probe (results not shown).

3.2.3. Initial Identification of Vicilin Clones by Southern Hybridisations.

Plasmid DNA was prepared from a number of clones which hybridised strongly to poly(A)$^+$ mRNA but not to the pDUB3 (legumin) cDNA probe. The plasmid mini-preps were digested with BamHI to excise their cDNA inserts, fractionated on agarose gels, and blotted onto nitrocellulose filters. The blots were probed with the labelled vicilin cDNAs from pDUB2 and pDUB4 (Lycett *et al.*,1983a) and washed under high stringency conditions. Fig.5A shows a 1% agarose gel of restricted, plasmid samples and the corresponding autoradiograph of the blotted DNA probed with the pDUB4 insert is shown in Fig.5B. After exposure of the autoradiograph, the nitrocellulose filter was washed for three 20 min.periods in H_2O at 65°C to remove the hybridised probe. The filter was then

FIGURE 5. Southern blot hybridisations of vicilin cDNA probes to recombinant plasmids.

A: BamHI-restricted DNA samples (except track 1) were electrophoresed through a 1% agarose gel as follows:

1) λNM258 DNA restricted with HindIII;
2) pAD2-1;
3) pAD3-4;
4) pAD5-4;
5) pAD6-11;
6) pAD7-13;
7) pAD7-8;
8) pAD7-4;
9) pAD7-7;
10) pDUB2;
11) pDUB3;
12) pDUB4(insert indicated by an arrow).

B: The DNA was blotted onto a nitrocellulose filter and probed with the ^{32}P-labelled pDUB4 insert. The pDUB2 hybridisation band is indicated by an arrow.

C: The first probe was washed off and the filter was rehybridised to the ^{32}P-labelled pDUB2 insert. The pDUB4 hybridisation band is indicated by an arrow.

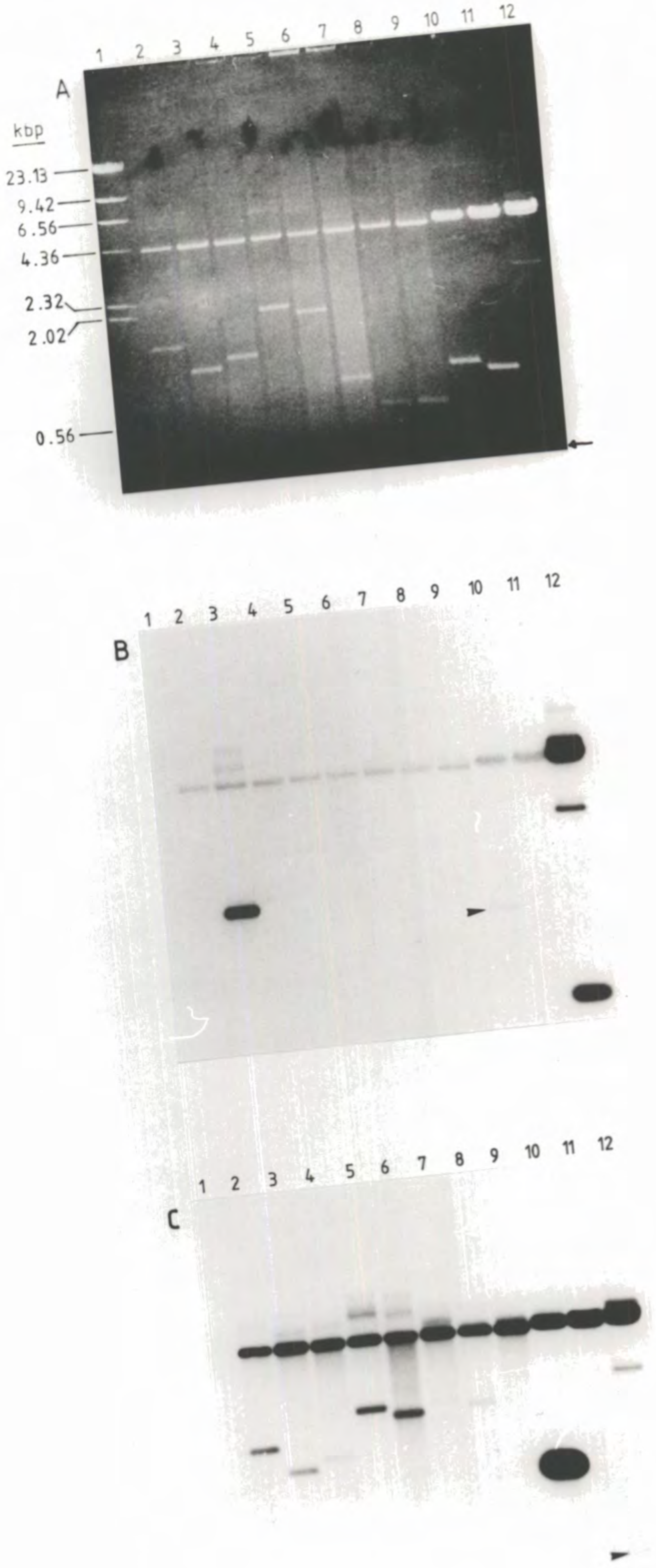

A
kbp
23.13
9.42
6.56
4.36
2.32
2.02
0.56
1 2 3 4 5 6 7 8 9 10 11 12
B
1 2 3 4 5 6 7 8 9 10 11 12
C
1 2 3 4 5 6 7 8 9 10 11 12

Fig. 5

reprobed with the pDUB2 excised insert (Fig. 5C). A conspicuous feature of the hybridisation results is that both the pDUB2 and pDUB4 insert probes, particularly the former, hybridised strongly to the vector (pBR322) DNA — the reason for this behaviour is not known. In addition to a restricted pDUB4 sample included as a control on the blot, only the pAD3-4 insert hybridised appreciably to the pDUB4 insert probe (Fig.5B). Weak hybridisations of that probe to the inserts from pAD2-1, 5-4, 6-11 and 7-13 were just discernible. The inserts from pAD2-1, 3-4, 6-11 and 7-13 hybridised to a significant extent to the pDUB2 insert probe though not nearly as strongly as that cDNA hybridised to itself (Fig.5C). pAD5-4, 7-4 and 7-7 hybridised weakly to the pDUB4 probe under high stringency conditions. Similar Southern hybridisation analyses (not shown) revealed that the cDNA inserts from pAD5-5, 8-6 and 8-14 hybridised weakly to both the pDUB2 and pDUB4 cDNAs while the pAD7-3 insert hybridised weakly to the pDUB2 insert only. Three plasmids, pAD5-8, 12-2 and 12-4 which contained single BamHI sites and were linearised upon digestion with BamHI were identified as vicilin clones by their weak hybridisation to the pDUB2 insert probe, but the high background hybridisation of the probe to the vector made that identification somewhat tentative. It is noteworthy that the inserts from pDUB2 and 4 did not cross-hybridise appreciably (Figs. 5B and 5C) although their sequences were ∿95% homologous (Lycett *et al.*,1983a). The sizes of the vicilin cDNA inserts ranged from ∿530 to 1950 bp (see Table 7.)

3.2.4. Southern Hybridisation Analysis of Putative Legumin Clones.

The presence of legumin cDNA inserts in legumin clones identified by colony screening with the pDUB3 cDNA probe was confirmed by Southern blot hybridisations. Most of the plasmids contained BamHI-excisable inserts which hybridised strongly to the ^{32}P-labelled pDUB3 cDNA. Apart from pAD6-15 which had an insert of <300 bp, the sizes of the legumin inserts ranged from ∿830 to ∿1100 bp (see Table 7). A remarkably high proportion of the plasmids, 13 out of 23, contained inserts of ∿830 bp. Some plasmids gave anomalous BamHI cleavage patterns : pAD 3-2 gave two fragments of ∿5200 and 1650 bp (note pBR322 = 4363 bp) both of which hybridised relatively weakly to the pDUB3 cDNA probe;

pAD's 4-12, 6-8 and 8-5 gave three BamHI fragments of ∿4400, 1000 and 530 bp of which only the 1000 bp fragment hybridised to the pDUB3 insert. The structures of these anomalous plasmids were not investigated further.

3.2.5. Restriction Mapping of Clones Isolated from the cDNA Bank.

pAD4-4 was chosen for further characterisation since it contained the longest insert (∿1100bp) among the pDUB3-hybridising clones. pAD10-5 was also selected due to its strong hybridisation to a 5' terminal fragment from pAD4-4 (see Section 3.2.2.). Restriction maps for the pAD4-4 and pAD10-5 legumin cDNA inserts are shown in Fig.6. Subsequent sequencing of these cDNAs revealed that the mapped restriction sites were accurate to within ± 30bp and the positions of the sites indicated in Fig.6 have been adjusted slightly to make them fully compatible with the sequence data. The restriction map of the previously sequenced insert from pDUB3 (Croy *et al.*, 1982) is presented for comparison with the pAD4-4 and 10-5 inserts. This comparison shows that the pAD4-4 cDNA extended beyond the 5'end of the pDUB3 insert by ∿300 bp, and that the pAD10-5 insert extended about another 400 bp towards the 5' end of the legumin mRNA but lacked a substantial portion (∿600 bp) of the 3' region. The alignment of common restriction sites among the three cDNAs showed that their overlapping segments were very similar except in a region at the 5' terminus of the pDUB3 cDNA which differed markedly from the corresponding regions in pAD4-4 and pAD10-5 —— pDUB3 contained an AvaI site ∿30 bp upstream from a BglI site common to all three cDNAs; pAD4-4 and pAD10-5 lacked that particular AvaI site but contained one ∿170 bp upstream from the BglI site. Since the pDUB3 and pAD4-4 cDNAs contained a second AvaI site in identical positions near their 3' termini, the two types of insert could be distinguished simply by the sizes of their respective AvaI internal fragments. Clones which had previously been shown to have inserts of the same length, ∿830bp, as pDUB3 (pAD1-4, 1-5, 7-8, 7-12, 9-3, 10-9, 11-3, 11-4, 11-5, 11-9, 12-1, 12-3, and 12-5) were restricted with AvaI and analysed by agarose gel electrophoresis. The separation of the AvaI sites within the inserts was found to be identical to that in the pDUB3 insert (results not shown). The orientations of the inserts in the vector were also the same as in pDUB3 except pAD9-3 which contained the insert in the opposite orientation.

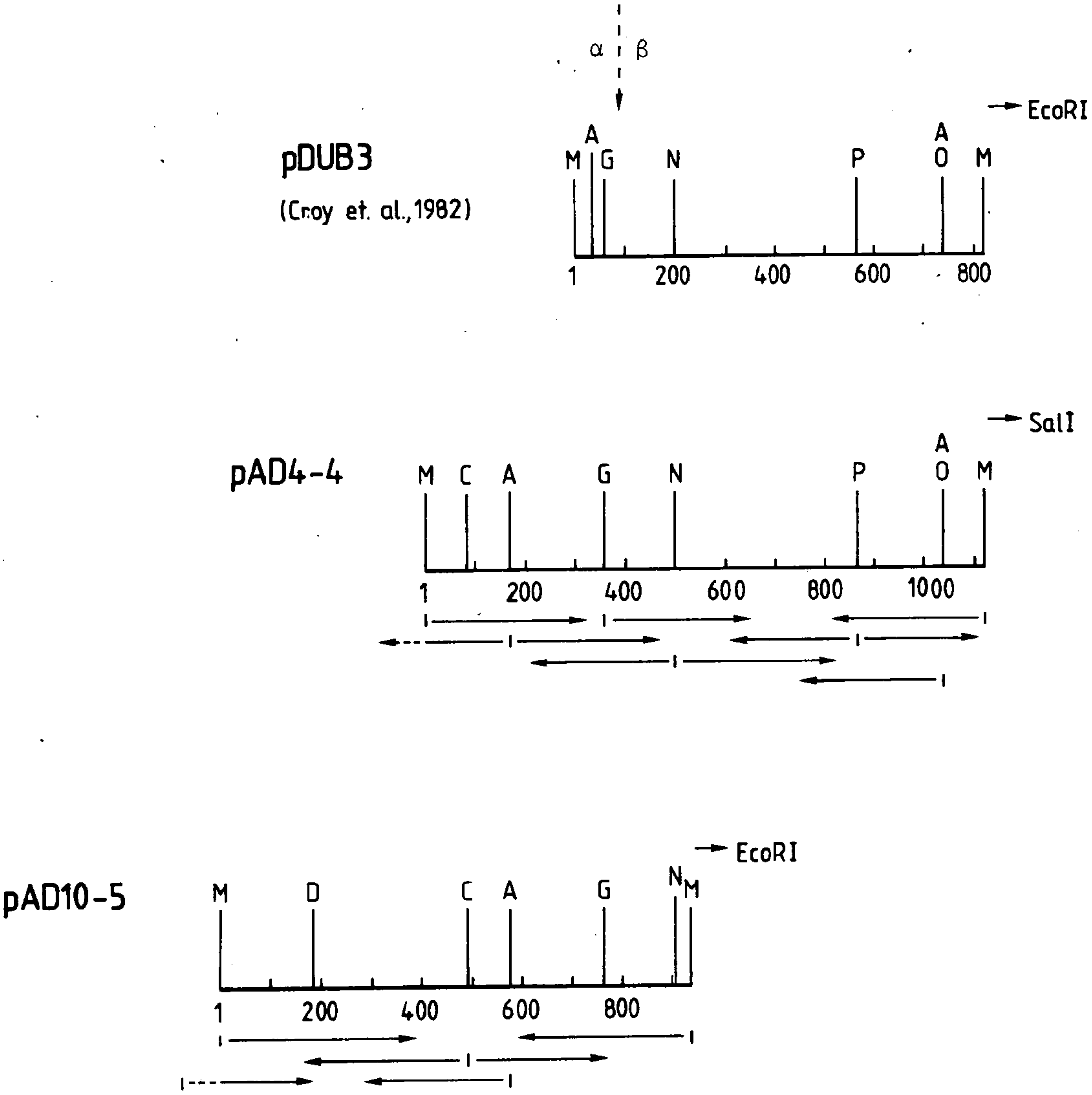


FIGURE 6. Restriction maps and sequencing strategy for various legumin cDNAs. The horizontal scales represent bp numbered from the 5' end of the coding strands. Solid arrows indicate the direction and extent of sequence determinations. The dashed arrow indicates vector sequences and is not drawn to scale.→EcoRI and →SalI indicate the orientation of the inserts relative to the EcoRI and SalI sites in pBR322. The BamHI sites at the termini of the inserts are linker sequences.

Restriction sites are abbreviated as follows :

A = AvaI; C = AccI; D = HindII; G = BglI; M = BamHI;
N = BstNI; O = XhoI; P = PstI.

Fig.7 shows a comparison of restriction maps for various vicilin cDNAs. The positions of restriction sites in the subsequently sequenced cDNAs have been adjusted to make them fully compatible with the sequence data. The maps of the pDUB4 and pAD3-4 inserts appeared to overlap, consistent with the strong hybridisation observed between these two cDNAs. Substantial regions of the pAD2-1, 6-11 and 7-13 maps were also very similar to one another but apart from these similarities, the different vicilin clones in general appeared to have very heterologous restriction maps. Thus, the cDNAs in Fig.7 were aligned mainly on the basis of the positions of the α : β and β : γ processing sites in the inserts (predicted from the DNA sequences, see Fig.11), but in the absence of sequence data, the maps of pAD5-4 and pAD7-3 were tentatively aligned on the basis of common restriction sites. Preliminary restriction mapping of the inserts from pAD5-5 and pAD7-4 (results not shown) indicated that both these cDNAs were distinct from the other cDNAs shown in Fig.7.

A restriction map for the BamHI insert (1850 bp) from pAD9-2 (which hybridised strongly to poly(A)$^+$ mRNA but not to legumin or vicilin cDNAs) is shown in Fig. 8. Preliminary restriction mapping data indicated that the insert from pAD6-2 was very similar to the pAD9-2 cDNA.

3.2.6. Characterisation of Vicilin cDNAs by Translation of Hybrid-selected mRNAs.

Clone pAD7-13, the insert from pAD3-4 and a pBR322 control were subjected to hybrid-selected translation (Fig.9). A degree of non-specific selection of abundant mRNAs by the pBR322 control was visible on the fluorograph. Cross-selection of two size classes of vicilin mRNAs by the filter-bound cDNAs was also evident which introduced some ambiguity in the interpretation of the results. Relative intensities of the hybrid-selected translation products suggested that the pAD3-4 insert predominantly selected mRNAs for a 47000-*M*r vicilin subunit, whereas pAD7-13 predominantly selected transcripts encoding a 50000-*M*r subunit.

3.2.7. Sequence Analysis of Legumin cDNA Clones.

Restriction maps illustrating the sequencing strategies for the pAD4-4 (1120 bp) and pAD10-5 (937 bp) legumin cDNA inserts are

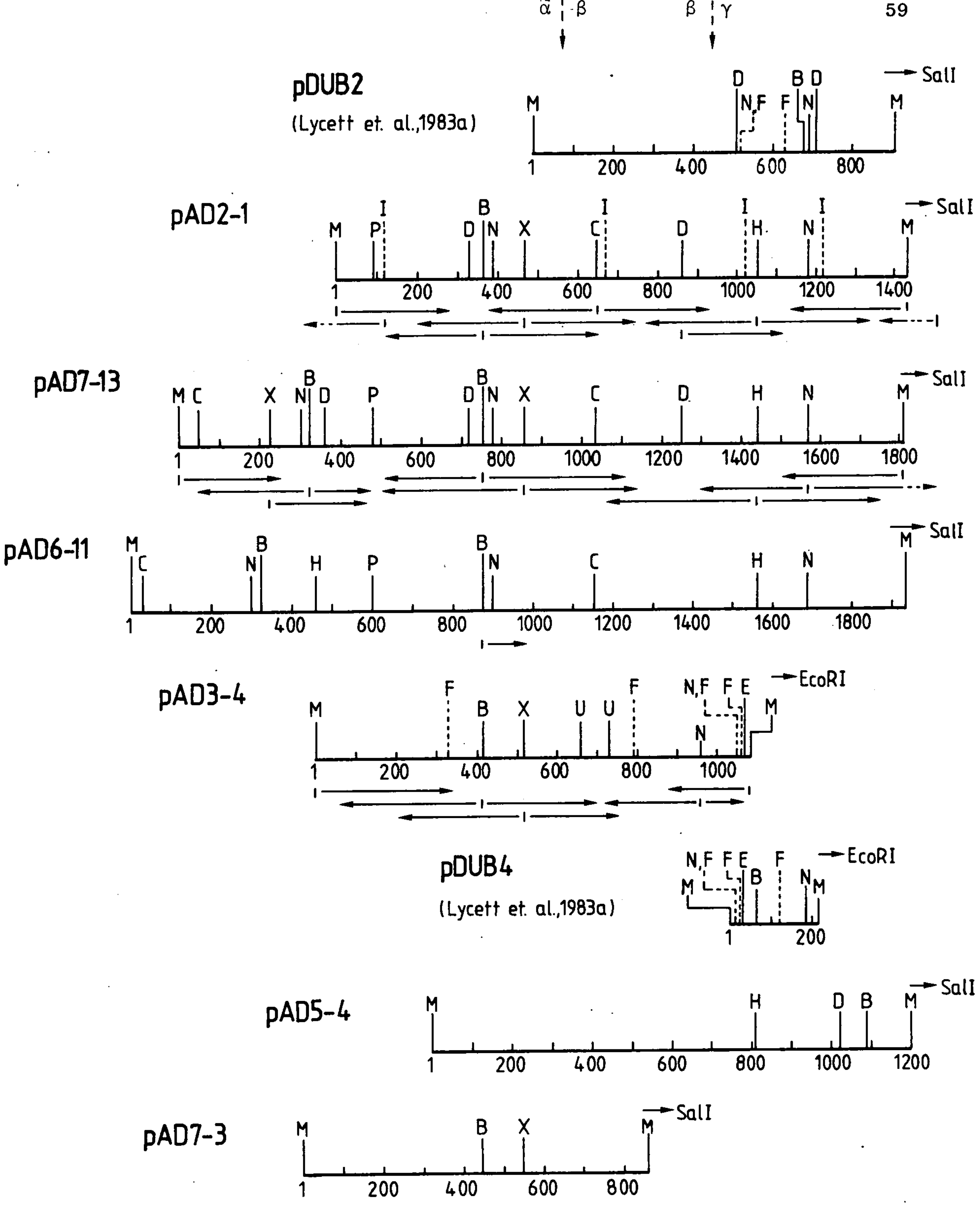


FIGURE 7. Restriction maps and sequencing strategies for various vicilin cDNAs. Symbols are as in Fig.6. with the following additions :

B = BglII; E = BstEII; F = HphI; H = HindIII; I = HinfI; U = Sau96I; X = XbaI.

The HinfI and HphI sites (indicated by dotted lines) were derived from the DNA sequences and are included only because these sites were used in sequencing or in the construction of hybrid vicilin cDNAs (see section 3.4.5.).

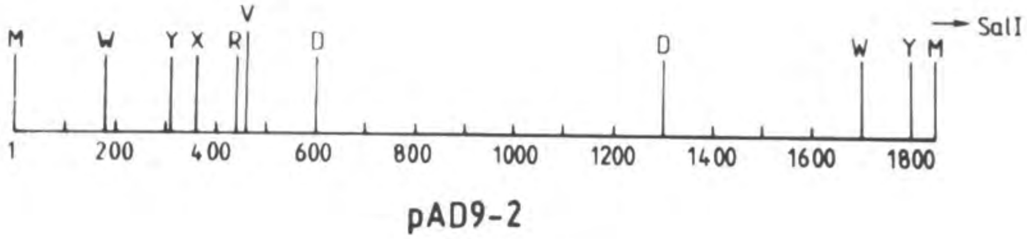

M
W
Y
X
R
V
D
D
W
Y
M
→ SalI
1
200
400
600
800
1000
1200
1400
1600
1800
pAD9-2

Fig. 8

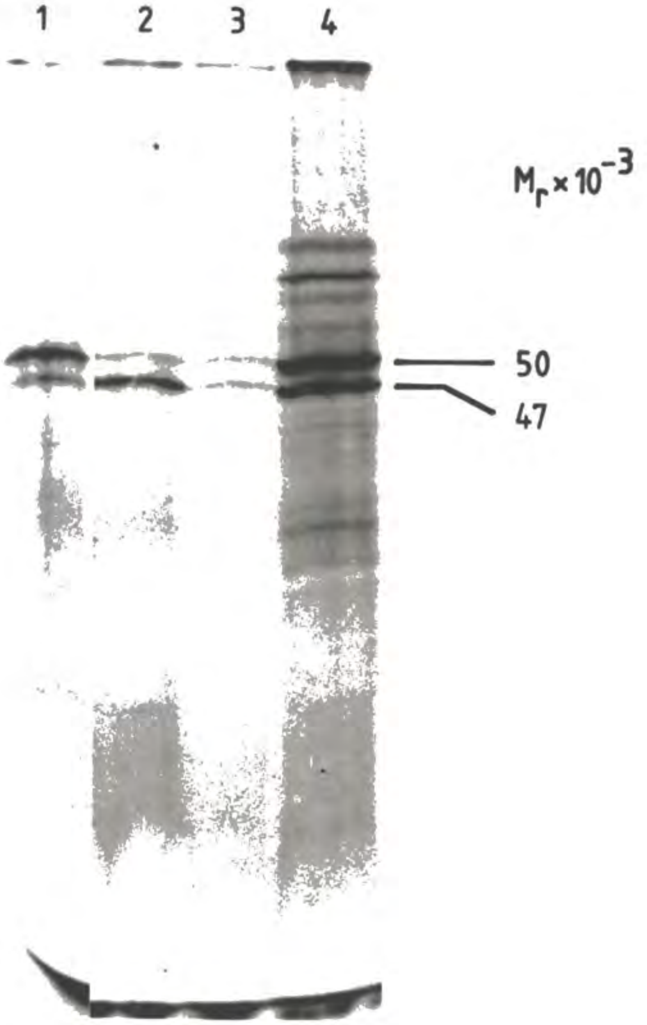

1
2
3
4
$M_r \times 10^{-3}$
50
47

Fig. 9

shown in Fig.6. Fig.10 shows the sequences of these cDNAs with the previously published composite sequence of pDUB1 and pDUB3 for comparison. The most conspicuous feature of the pAD4-4 and pAD10-5 sequences is the presence of three ∿54 bp tandem repeats in the region coding for the legumin acidic subunit. Only half of one repeat is present in pDUB3. Apart from these repeats, the legumin cDNA sequences share extensive (∿99%) homology.

3.2.8. Sequence Analysis of Vicilin cDNA Clones.

The cDNA inserts in pAD2-1, 3-4 and 7-13 were completely sequenced whereas only a short region of pAD6-11 was sequenced. Fig.7 shows the sequencing strategies for these vicilin cDNAs. The DNA sequences of pAD2-1 and 3-4 are shown in Fig.11 with the previously published sequences of the inserts from the vicilin clones pDUB2 and pDUB4 (Lycett *et al.*,1983a) for comparison.

There is extensive homology among the sequences of all four vicilin clones. In fact, the overlapping regions (52 bp) of pAD3-4 and pDUB4 are identical which suggests that the two cDNAs were derived from mRNA transcripts from the same gene, and in the following comparisons of sequence homology, reference to the pAD3-4 sequence means the composite pAD3-4/pDUB4 sequence. Where sequence data are available for pairwise comparisons to be made, the coding regions of pAD2-1 and pAD3-4 are 83.9% homologous, pAD2-1 and pDUB2 are 85.6% homologous, which pAD3-4 and pDUB2 are 83.9% homologous. The 3' noncoding regions of the cloned cDNAs show more variation than the coding sequences. There are 18 out of 73 (25%) bp mismatches in the pDUB2 and pAD2-1 sequences following the doublet stop codons; the pAD3-4 sequence does not extend into this region to allow sequence.comparisons.

Though not included in Fig.11, the sequence of another vicilin cDNA, pAD7-13, was determined (see Fig.7). The pAD7-13 sequence was very similar to the pAD2-1 sequence over an internal region of ∿1360 bp. Only 2 bp differences occured within that overlapping region —— these are indicated in Fig.11. However, the pAD7-13 sequence diverged significantly from the pAD2-1 sequence at both ends of the inserts. The relationship between pAD7-13 and pAD2-1 is shown schematically in Fig.12. At the end

```
pAD10-5   G GCC CTC TCT CGT GCT ACC CTT CAA CGC AAC GCC CTT CGC AGA CCT TAC TAC TCC AAT GCT CCC CAA GAA ATT TTC ATC CAA CAA GGT

pAD10-5 AAT GGA TAT TTT GGC ATG GTA TTC CCC GGT TGT CCT GAG ACC TTT GAA GAG CCA CAA GAA TCT GAA CAA GGA GAG GGA CGC AGG TAC AGA

pAD10-5 GAC AGA CAT CAA AAG GTT AAC CGA TTC AGA GAG GGT GAT ATC ATT GCA GTT CCT ACT GGT ATT GTA TTT TGG ATG TAC AAC GAC CAA GAC

pAD10-5 ACT CCA GTT ATT GCC GTC TCT CTT ACT GAC ATT AGA AGC TCC AAT AAC CAG CTT GAT CAG ATG CCT AGG AGA TTC TAT CTT GCT GGG AAC

pAD10-5 CAC GAG CAA GAG TTT CTA CAA TAC CAG CAT CAA CAA GGA GGA AAG CAA GAA CAA GAA AAT GAA GGC AAC AAC ATT TTC AGT GGC TTC AAG
pAD4-4                                                                   .. ... ... ... ... ... ... ... ... ... ... ... ... ...

pAD10-5 AGG GAT TAC TTG GAA GAT GCT TTC AAC GTG AAC AGG CAT ATA GTA GAC AGA CTT CAA GGC AGG AAT GAA GAC GAA GAG AAG GGA GCC ATT
pAD4-4  ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ...

                                                    AvaI ↓ |←———————————————————————————— Repeat 1 ———————————————————————
pAD10-5 GTC AAA GTG AAA GGT GGA CTC AGC ATC ATA AGC CCA CCC GAG AAG CAA GCG CGC CAC CAG AGA GGC AGC AGA CAA GAG GAA GAT GAA GAT
pAD4-4  ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... .A. ... ... ... ... ... ...
pDUB3            .. ... ... ... ... ..G ... ... ... ... ... ... ... ... ... ... ... ... ... ... --- --- --- --- --- --- --- ---

         →|←————————————————————————————————————— Repeat 2-(       )——————————————————————→|←————————————————— Repeat 3 ———————
pAD10-5 GAA GAG AAG CAG CCG CGC CAC CAG AGA GGC AGC AGA CAA GAG GAA GAG GAA GAT GAA GAT GAA GAG AGG CAG CCG CGT CAT CAA AGG AGA
pAD4-4  ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ...
pDUB3   --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- ---

        ——————————————————————————————————→|                            ↓BglI                            α     ↓     β
pAD10-5 AGA GGA GAG GAG GAA GAA GAA GAC AAG AAA GAG CGC GGC GGC AGC CAA AAA CGC AAA AGC AGA AGG CAA GGA GAC AAT GGG CTT GAG GAA
pAD4-4  ... ... ... ... ... ... ... ... ... ... ... ... C.. ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ...
pDUB3   --- --- --- --- --- --- --- --- --- --- --- --- --- --- ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ...

pAD10-5 ACA GTT TGC ACT GCT AAA CTT CGA TTG AAC ATT GGC CCG TCT TCA TCA CCA GAC ATC TAC AAC CCT GAA GCT GGT AGA ATC AAA ACT GTT
pAD4-4  ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ...
pDUB3   ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ...

pAD10-5 ACC AGC CTG GAC CTC CCA GTT CTC AGG TGG CTC AAA CT
pAD4-4  ... ... ... ... ... ... ... ... ... ... ... ... ..A AGT GCT GAG CAT GGA TCT CTC CAC AAA AAT GCT ATG TTT GTG CCT CAC TAC
pDUB3   ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... A.. ... ... ..A ... ... ...

pAD4-4  AAC CTG AAT GCA AAC AGT ATA ATA TAC GCA TTG AAG GGA CGT GCA AGG CTA CAA GTA GTG AAC TGC AAT GGC AAC ACC GTG TTT GAT GGA
pDUB3   ... ... ... ... ... ..C ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ...

pAD4-4  AAG CTA GAA GCC GGA CGT GCA TTG ACA GTG CCA CAA AAC TAT GCT GTG GCT GCA AAG TCA CTA AGC GAC AGG TTC TCA TAT GTA GCA TTC
pDUB3   ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... .A. ... ... ... ... ... ... ... ...

pAD4-4  AAG ACC AAT GAT AGA GCT GGT ATT GCA AGA CTT GCA GGG ACA TCA TCA GTT ATA AAT AAT CTG CCG TTG GAT GTG GTT GCA GCT ACA TTC
pDUB3   ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... G.. ... ..C ... ... ... ... ... ... ... ...

pAD4-4  AAC CTG CAG AGG AAT GAG GCA AGG CAG CTC AAG TCC AAC AAT CCC TTC AAA TTT CTA GTT CCA GCT CGT CAG TCT GAG AAC AGA GCT TCG
pDUB3   ..A ... ... ... G.. ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ...

pAD4-4  GCT TAG att tcg cac caa atc aat gaa agt aat gaa taa gaa aac taa ggc tta gat gcc ttt gtt act tgt gta aaa taa ctc gag tca
pDUB3   ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ...

pAD4-4  tgt acc ttt ttg cgg aaa cag aat aaa taa aag gta aaa ttt cag tgc tct aaa aaa aaa aaa aaa aaa aaa aaa aaa aaa aaa aaa a
pDUB3   ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... .
```

Figure 10. Comparisons of the nucleotide sequences of the cDNA inserts from pAD10-5 and pAD4-4, and the previously published composite sequence of pDUB3/pDUB1 (Croy *et al.*, 1982). Dots represent nucleotides which are identical to those in the uppermost sequence. Broken lines (- - -) indicate gaps inserted in the sequences to maximize homology. The repeats are indicated by labelled arrows over the pAD10-5 sequence, and the hexanucleotide duplication in the second repeat (GAGGAA) is bracketed. The cleavage site between the α- and β-legumin subunits is indicated by a vertical arrow. The 3' noncoding sequences are shown in lower case, and the consensus polyadenylation signal sequence, AATAAA, is boxed (□). Other putative signal sequence variants are enclosed in broken boxes (⬚). Restriction sites used for the preparation of specific legumin cDNA probes (see section 3.3.) are labelled.

```
pAD3-4   CA ATC AAA CCG TTA ATG TTG TTG GCA ATT GCT TTC CTA GCC TCA GTT TGT GTC TCT TCT AGA TCC GAT CAA GAG AAC CCC TTT ATC TTT
pAD2-1                                                                       .T ... ... ..G ..T ... .CT C.A ..T ..T ... ... ..C

pAD3-4  AAG TCT AAC CGA TTT CAA ACT CTT TAT GAG AAC GAA AAC GGT CAC ATT CGT CTT CTC CAA AAA TTT GAC AAA CGT TCC AAA ATA TTT GAA
pAD2-1  ... ... ... AAG ... ... ... ... .T. ... ..T ... ..T ..G ... ... ..A ... ..G ..G ... ... ... C.. ... ..T ... ..T ..C ..G
             *

pAD3-4  AAT CTT CAA AAT TAC CGT CTT TTA GAA TAT AAG TCC AAA CCT CAC ACC CTT TTT CTT CCA CAA TAC ACC GAT GCC GAC TTC ATC CTT GTA
pAD2-1  ... ..A ... ..C ... ... ... ..G ... ... ... ... ... ... ... ..A A.A ... ... ... ..G C.. ... ... ... ..T .A. ... ... ..T

pAD3-4  GTC CTT AGT GGG AAA GCC ACT CTC ACA GTG TTG AAA TCT AAC GAT CGA AAC TCC TTC AAT CTT GAA CGT GGT GAT GCC ATC AAA CTC CCT
pAD2-1  ..A ..C ... ..A ... ..T .TA ... ... ... ... ... C.C G.T ... A.. ... ... ... ..C ... ..G ..C ..A ... A.G ..A ... ..T ...

pAD3-4  GCT GGC TCT ATT GCT TAT TTC GCT AAC CGA GAT GAC AAC GAG GAA CCT AGA GTA TTA GAT CTC GCC ATC CCA GTA AAC AAA CCC GGT CAA
pAD2-1  ... ... A.A ... ... ... ..G .T. ... A.. ... ... ... ... ..G .T. ... ... ... ... ... ... ..T ..C ... ..T .G. ..T ..C ...

pAD3-4  TTG CAG TCT TTC TTA TTA TCT GGA ACT CAA AAT CAA AAA TCA TCA TTA TCT GGA TTC AGC AAG AAT ATT CTA GAG GCT GCT TTC AAT ACC

pAD2-1  C.T ... ... ... ... ..G ... ... .A. ... ..C ... C.. AAC .AC ... ... ..G ... ..T ... ..C ... ... ... ... T.C ... ... ..T
                                                                                                         α ↓ β
pAD3-4  AAT TAC GAG GAG ATA GAA AAG GTT CTT TTA GAA CAA CAG GAA CAA GAG CCA CAA CAC AGA AGA AGT CTT AAG GAT AGG AGA CAA GAG ATC
pAD2-1  G.T ..T ..A ... ... ... ... ... ... ... ... G.G ..T ..G A.. ... A.. ... ... ... ... ..C ... ... ... .A. ..G ..G C.A .GT
pDUB2   G.C A.T .CA ... ... ..G ... A.. ..C ... ... G.G ..T ..G A.. ... A.. ..T ... ... ... G.C ... .G. ... .A. ... ... C.. .G.

pAD3-4  AAC GAA GAA AAT GTA ATA GTC AAA GTA TCA AGG GAC CAA ATT GAG GAA TTG AGC AAA AAT GCA AAG TCC AGT TCC AAA AAA AGT GTA TCA
pAD2-1  C.A ... ..G ... ... ... ..A ... T.. ... ... .GA ... ... ... ... ... ..T ... ... ... ... ..T .CC ... ... ... ... ..T ..C
                                                                                                                     G
pDUB2   C.A ... A.G ... ... ... ... ... ... ... ..G A.A ... ... ..A ... ..A ..T ... ..C ... ... ..T ..C ... ... ... ... ... ..T

pAD3-4  TCA GAA TCT GGA CCA TTC AAC TTG AGA AGT CGG AAT CCT ATC TAT TCT AAC AAG TTT GGC AAA TTC TTT GAG ATC ACC CCA GAG AAA AAT
pAD2-1  ..T ... ... .A. ... ... ... ... ... ... ..C GG. ... ... ... ..C ... G.. ... ..A ... ... ... ..A ... ... ... ... ... ...
pDUB2   ..T CG. ..A .A. ... ... ... ... .A. ... A.T G.. ... ... ... ..C ... C.A .A. ..T ... ... ... ... ... ..T ... A.. ... ...

pAD3-4  CAA CAA CTT CAA GAC TTG GAC ATA TTT GTC AAT TCT GTG GAT ATT AAG GAG GGA TCT TTA TTG TTG CCA AAC TAC AAT TCA AGA GCA ATT
pAD2-1  .C. ..G ... ... ... ... ..T ... ... ... ... ... ..A ..G ... ... ... ... ... ... ... ... ... C.. ... ... ... ..G ..C ..A
                                                                                    C
pDUB2   .C. ... ... ... ... ... ..T ... ... ... ... .A. ... ..G ... ..A ... ..G ... C.. .G. ... ... C.. ..T ... ... ..G ..C ..A
                                                                                                                         β ↓ γ
pAD3-4  GTG ATA GTA ACT GTT ACC GAA GGA AAA GGA GAT TTT GAA CTT GTG GGT CAA AGA AAT GAG AAC CAG --- --- --- GGA AAA GAA AAT GAC
pAD2-1  ..A ... ... ..A ... .A. ... ... ... ... ... ... ... ... ... ... ... ... ... ..A ... ..A CAA GAG CAG A.. ... ... G.. ...
pDUB2   ... ... ... ..A ... .AT ... ... ... ..G ..C ... ... ... ... ... ... ... ... ... ... ..A CAA GGC TTG A.. G.. ... G.. ...

pAD3-4  AAG GAA GAG GAA CAA --- GAA GAA GAG ACA AGC AAA CAA GTG CAA CTG TAT AGA GCT AAG TTG TCT CCA GGT GAT GTT TTT GTG ATT CCA
pDUB4                                                                               ... ... ... ... ... ... ... ... ... ... ...
pAD2-1  G.. ... ... ... ... GGA ... ..G ... .T. .AT ... ... ... ... AAT ..C .A. ... ..A ..A ... T.. ..A ... ... ... ... ... ...
pDUB2   G.. ... ... ..G ... AGA ... ... ... ... .AG ..T ... ... ... AGT ..C .A. ... ... ... A.. ... ... ... ... ... ..A ... ..G

pAD3-4  GCA GGT CAC CCC GTT GCC A
pDUB4   ... ... ... ... ... ... .TA AAT GCC TCC TCA GAT CTC AAT CTG ATT GGA TTG GGT ATC AAT GCC GAG AAC AAC GAG AGA AAC TTC CTT
pAD2-1  ... ..C ..T ..A ... ... C.. ..A ..T ... ... A.. ..T G.. T.. C.. ..G ..T ... ..T ... ..T ... ... ..T C.. ..G ... ..T ...
pDUB2   ... ... ..T ..T ... ... G.. .GA ..T ..A ... A.. ... ... T.. C.. ... ..T ... ... ... ..T ..A ... ... ... ... ... ... ...

pDUB4   GCA GGT GAG GAA GAC AAT GTC ATA AGT CAA GTA GAA AGA CCA GTT AAA GAG CTT GCA TTT CCT GGA TCT TCT CAT GAG GTT GAT AGG
pAD2-1  ... ..C ..T ..G ..T ... ..G ..T ... ..G A.. C.G C.. ... ..G ... ... ... ... ..C ... ... ..A G.. ..A ... ... ... ... ATA
pDUB2   ... ... ... ..G ... ... ..G ... ... ..G A.. C.G .A. .A. ..G ... ..T ... A.. ... ... ... ... G.. ..A ... ... ..C ... C..

pAD2-1  CTA GAG AAT CAG AAA CAA TCC CAC TTT GCA GAT GCT CAA CCT CAA CAA AGG GAG AGA GGA AGT CGT GAA ACA AGA GAT CGT CTA TCT TCA
pDUB2   ... ... ... ..A ... ... ..T T.T ... ... A.. ... ... ... ... ... ..A ... .C. A.. ..C .AA ... .T. .AG ..A .A. ..G .A. ...

pAD2-1  GTT TGA aat gtt tct taa tga gtg gac aaa ata cta tgt atg tat gct atc aag aga tat atc tca cgg gga gga atg aat aaa aca atg
pDUB2   A.. .TG GGG .CC .T. ... ... .a. at. ... ... t.t ..c ... ... ... ..a ... .ac ... .g. ... taa t.. ... .g. ... ... ... tc.

pAD2-1  tta tct tat aac tat aat tat ata tcc act ttt cta cta tga ata
                         *

pDUB2   ..c ... .
```

Figure 11. Comparisons of the nucleotide sequences of the cDNA inserts from pAD3-4 and pAD2-1, and the previously published pDUB2 and pDUB4 sequences (Lycett *et al.*, 1983a). The two asterisks (*) below the pAD2-1 sequence define the boundaries of the sequence which is homologous to the pAD7-13 sequence. The 430bp sequence upstream from the first asterisk in pAD7-13 comprises an inverse repeat (see Fig.12), whereas a poly(A) tail extends downstream from the second asterisk. Where pairs of nucleotides are shown in the pAD2-1 sequence, the lower one indicates the nucleotide found at that position in pAD7-13. The codon specifying the N-terminus of a mature 50000-*M*r subunit is underlined. Other symbols are as used in Fig.10.

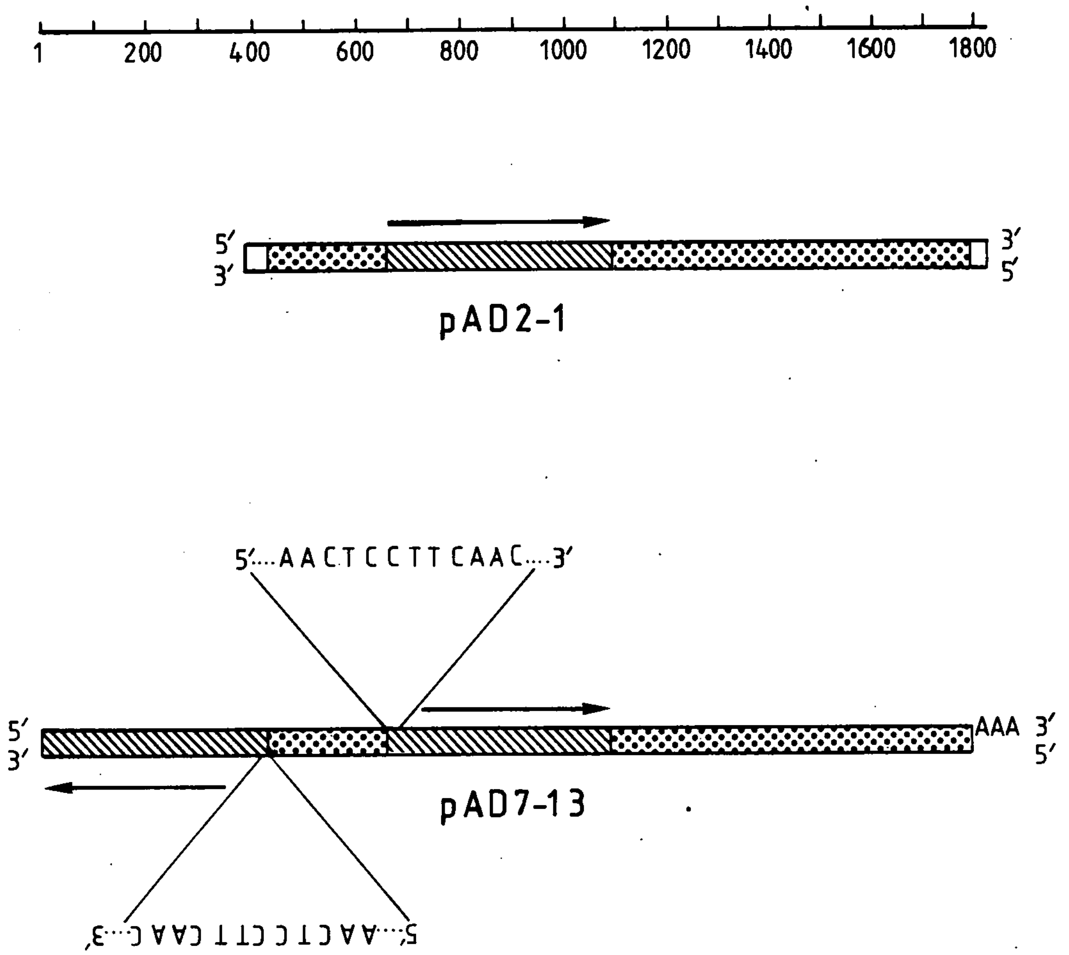


FIGURE 12: Schematic representation of the relationship between the pAD2-1 and pAD7-13 cDNAs. The horizontal scale is numbered in bp from the 5' end of the pAD7-13 coding strand.

░ = regions which are virtually identical in both clones.
▨ = 430bp inverted repeat at the 5' end of pAD7-13.
▨ = internal sequence, common to both clones, which is duplicated in pAD7-13.
□ = sequences unique to pAD2-1.

The DNA sequence at the 5' end of the duplicated region is indicated to illustrate the orientation of the repeat. AAA indicates the presence of a poly(A) tail in the pAD7-13 cDNA.

corresponding to the 5' terminus of the coding strand in pAD7-13 was a 430 bp sequence which was an exact inverse repeat of an internal sequence located 225 bp further downstream. The inverse repeat contained 7 in-phase stop codons and shared no homology with the corresponding regions in the pAD2-1 and pAD3-4 sequences. At the 3' ends of the inserts, pAD7-13 had a poly(A) tail attached to an A residue (indicated by an asterisk in Fig.11) 32bp upstream from the 3' terminus of the pAD2-1 cDNA coding strand.

Only a short region of pAD6-11 was sequenced (see Fig.7). The 90 nucleotide-long sequence obtained was identical to the sequences of the corresponding regions in pAD2-1 and pAD7-13 (see Fig.7).

3.3. Probing of Pea Genomic Digests with Legumin cDNA Probes.

Genomic DNA from pea leaves was digested to completion with various restriction enzymes as indicated in Fig.13, fractionated in three aliquots on a 1% agarose gel, and blotted onto a nitrocellulose filter. The filter was subsequently cut into three strips each containing a replicate sample of the resolved genomic fragments. The three samples were individually probed with ^{32}P-labelled fragments of legumin cDNAs corresponding approximately to the coding sequences for the basic subunit (pAD4-4 BglI - BamHI 765 bp fragment), the acidic repeats (pAD4-4 AvaI - BglI 193 bp fragment), and the acidic subunit upstream from the repeats (pAD10-5 BamHI - AvaI 580 bp fragment). The results are shown in Fig. 13. The patterns of the bands hybridising to the probes for the acidic and basic subunit regions were practically identical. All the fragments which hybridised to these two probes also hybridised to the probe for the repeat units though there was some variation in the relative intensities of the hybridisation bands. The latter probe also hybridised weakly, but distinctly, to a number of additional genomic fragments which did not bind the acidic and basic subunit probes at the stringency used (0.5 x SSC at 60°C).

3.4. Construction of Vicilin Expression Plasmids.

3.4.1. pAD2-1.expl.

The construction of pAD2-1.expl involved the ligation of the

FIGURE 13. Southern blot hybridisation of specific regions of the legumin cDNAs to pea genomic DNA. 10μg aliquots of pea DNA were restricted with EcoRI, HindIII, and HindIII/BamHI, and electrophoresed through a 1% agarose gel. The DNA was transferred to nitrocellulose paper, and probed with ^{32}P-labelled cDNA fragments (specific activity ≃10^8 cpm/μg) corresponding approximately to the legumin basic subunit (b), the acidic region upstream of the repeats (a), and the repeats in the acidic subunit (r). The top three autoradiographs were exposed for seven days, whereas the bottom three were exposed for 2 days. (N.B. A track containing pBR322 size markers was cut off from the autoradiographs since labelled, contaminating pBR322 sequences in the basics probe hybridised strongly to these fragments. The intense hybridisations have "spilled over" and are visible in tracks hybridised with the basics probe).

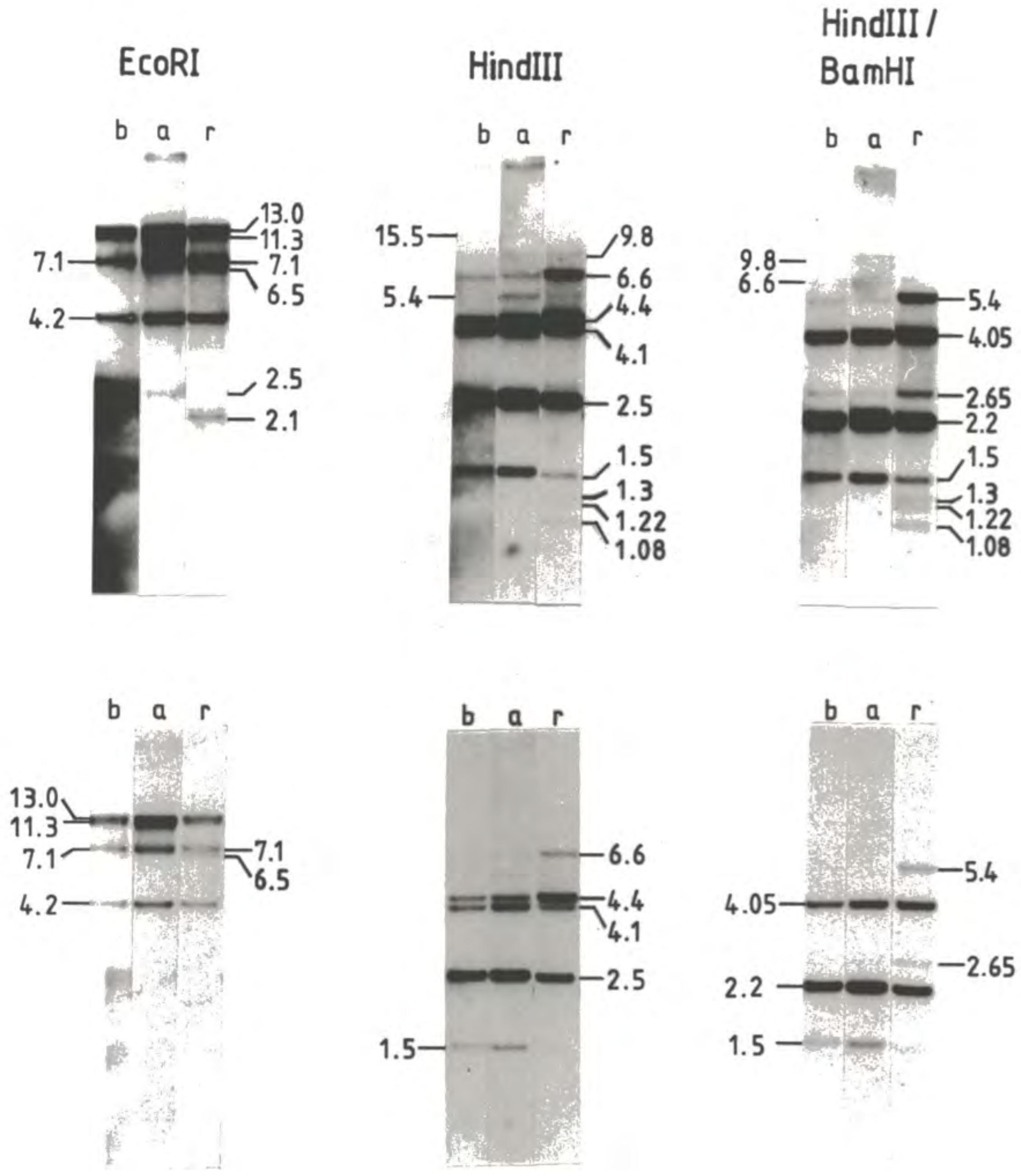

EcoRI
HindIII
HindIII / BamHI
b a r
b a r
b a r
13.0
11.3
7.1
6.5
7.1
4.2
2.5
2.1
15.5
5.4
9.8
6.6
4.4
4.1
2.5
1.5
1.3
1.22
1.08
9.8
6.6
5.4
4.05
2.65
2.2
1.5
1.3
1.22
1.08
b a r
b a r
b a r
13.0
11.3
7.1
4.2
7.1
6.5
6.6
4.4
4.1
2.5
1.5
5.4
4.05
2.65
2.2
1.5

Fig. 13

BamHI insert from pAD2-1 (see Fig.6.) directly into the BamHI site of the expression plasmid pPLc24 (see Section 2.1.3.). The construction is shown schematically in Fig.14. pPLc24 was linearised with BamHI, treated with alkaline phosphatase, and ligated in 3-fold molar excess to the BamHI insert from pAD2-1. In this ligation, the coding sequence of the pAD2-1 cDNA was inserted in the same reading frame as that of the MS2 replicase gene present in the vector. *E.coli* K12ΔH1Δ*trp* was transformed to ampicillin resistance with the ligation mixture. Out of 115 transformants screened by colony hybridisation with a ^{32}P-labelled pAD2-1 insert probe, 65 were positive. Plasmid DNA from 8 randomly chosen positive clones was digested with HindIII and run on an agarose gel to determine the orientation of their pAD2-1 inserts. Out of five plasmids with the insert in the appropriate orientation for expression, one was chosen for further work and was designated pAD2-1.expl(+). One of the 3 plasmids with the insert in the opposite orientation was designated pAD2-1.expl(-) and was used as a negative control.

3.4.2. pAD2-1.exp2.

For the construction of pAD2-1.exp2,the pAD2-1 cDNA insert was specifically "trimmed" so that the 5' terminal codon encoded the N-terminus of a mature vicilin subunit. The trimmed cDNA was then inserted by blunt-end ligation into the expression plasmid pPLc 245 (see Section 2.1.3.). The construction is shown schematically in Fig.15. Plasmid pPLc245 was linearised with SalI, made blunt-ended with mung-bean nuclease and then treated with alkaline phosphatase. Plasmid pAD2-1 was digested with BamHI and treated with T4 DNA polymerase in the presence of dTTP. In the T4 polymerase reaction, each DNA strand was degraded in a 3'→ 5' direction until the first A residue was encountered on the opposite strand, at which point the exonuclease activity of the enzyme was masked by its stronger 5' → 3' polymerase activity. In the pAD2-1 cDNA, the first A encountered from the 5' end of the coding strand happens to be the A of the N-terminal AGG codon of mature vicilin. Thus as a result of the T4 polymerase digestion, a 15 bp 5' protruding end was generated upstream from the N-terminal codon; this single-stranded extension was removed with mung-bean nuclease. The resulting blunt-ended fragment was ligated to a 5-fold molar

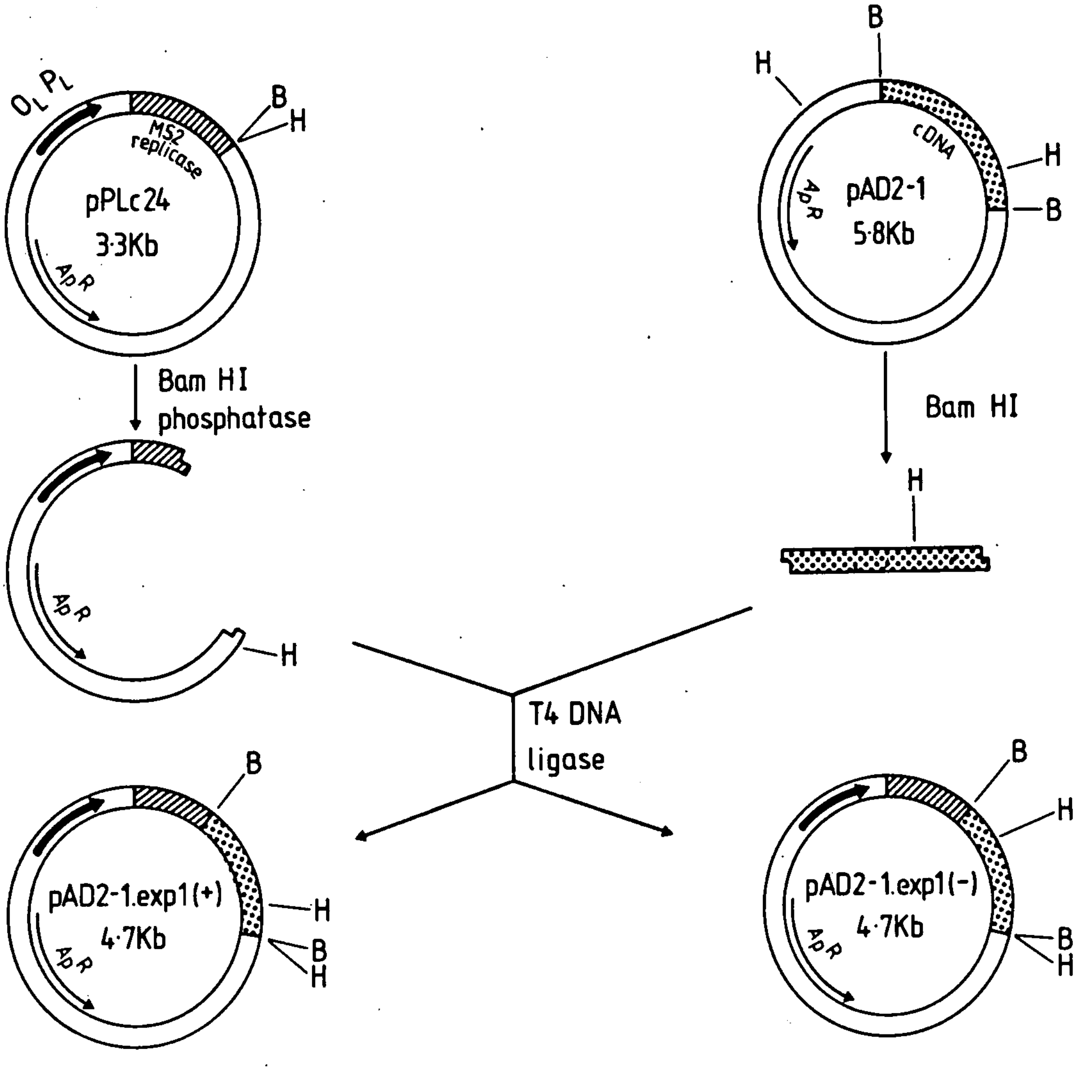


Figure 14. Construction of pAD2-1. exp1 (see text for details)

➡ = phage λ operator-promoter region showing direction of transcription.

▨ = ribosome binding site and sequence encoding the N-terminal 98 amino acids of the phage MS2 replicase gene

⬚ = vicilin cDNA insert in pAD2-1

Ap^R = ampicillin resistance gene

Only relevant restriction sites are shown and are abbreviated as follows :

B = BamHI; H = HindIII

pAD2-1.exp1(+) contains the cDNA insert in the appropriate orientation for expression.

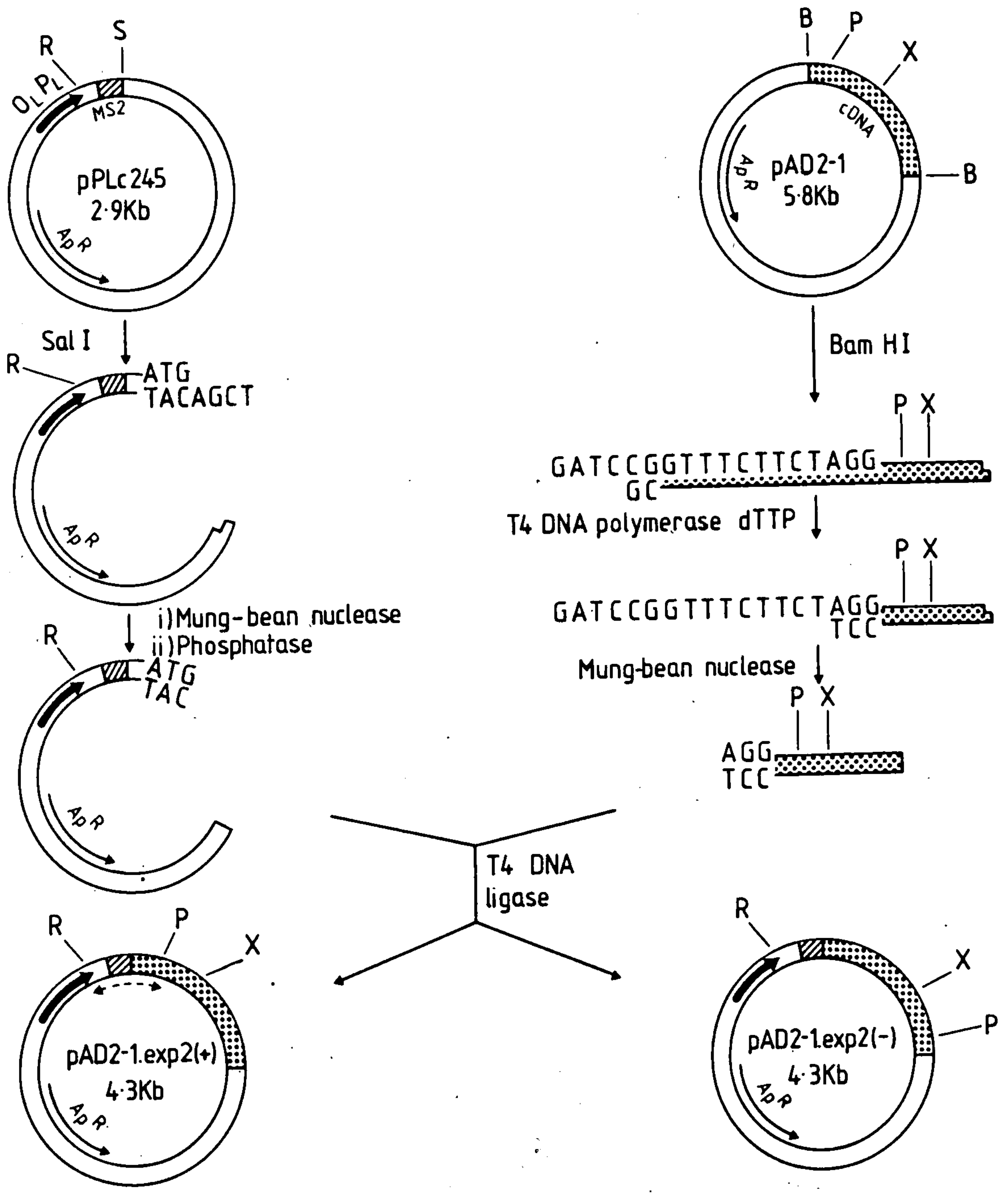


Figure 15. Construction of pAD2-1.exp2 (see text for details).

▨ = ribosome binding site of the phage MS2 replicase gene.

B = BamHI; P = PstI; R = EcoRI; S = SalI; X = XbaI

Other symbols are used in Fig. 14. The nucleotide sequence across the MS2-cDNA junction (←-→) was verified by DNA sequencing. pAD2-1.exp2(+) contains the cDNA insert in the appropriate orientation for expression.

excess of the mung-bean nuclease-treated pPLc245, and *E. coli* K12ΔH1Δ*trp* was transformed to ampicillin resistance with the ligation product. Out of 800 transformants screened by colony hybridisation with a ^{32}P-labelled pAD2-1 insert probe, 4 were positive. Plasmid DNA from these 4 clones was digested with EcoRI and XbaI to determine the orientations of their pAD2-1 inserts. One plasmid, designated pAD2-1.exp2(+) contained the insert in the correct orientation for expression. The three other plasmids, one designated pAD2-1.exp2(-), contained the insert in the opposite orientation. To verify the construction of pAD2-1.exp2(+), an EcoRI-PstI fragment spanning the junction between the expression vector and the 5' end of the pAD2-1 insert (see Fig.15) was sequenced. The sequence across the junction read:5'...AGGATTACCC<u>ATG</u>AGGTCT...3' (initiation codon underlined) which confirmed that the desired construction had been achieved (see Remaut *et al.*1983a). The remainder of the determined sequence was in complete agreement with the published sequence of the MS2 genome (Min Jou *et al.*1972) and the pAD2-1 sequence (Fig.11).

3.4.3. pAD2-1.exp3.

In the construction of pAD2-1.exp3, the BamHI insert from pAD2-1 was ligated directly into the BamHI site of pAS1 (see Section 2.1.3.). The construction is shown in Fig.16. The cDNA insert of plasmid pAD2-1 was excised with BamHI, treated with alkaline phosphatase, and ligated in ~10-fold molar excess to a sample of BamHI-linearised pAS1. *E. coli* N99λ*c*I857 was transformed to ampicillin resistance with the ligation products. Out of 26 transformants screened by colony hybridisation with a ^{32}P-labelled pAD2-1 insert probe, 13 were positive. Out of 8 positive plasmids restricted with HindIII and analysed by agarose gel electrophoresis, 4 contained the insert in the appropriate orientation for expression. One of these was designated pAD2-1.exp3(+). A plasmid with the insert in the opposite orientation was designated pAD2-1.exp3(-).

3.4.4. pAD2-1.exp4.

The strategy used for the construction of pAD2-1.exp4 was identical to that for the construction of pAD2-1.exp2 except that pAS1 (see Section 2.1.3.) was the expression plasmid used (see Fig. 16.). pAS1 was linearised with BamHI and treated with mung-bean

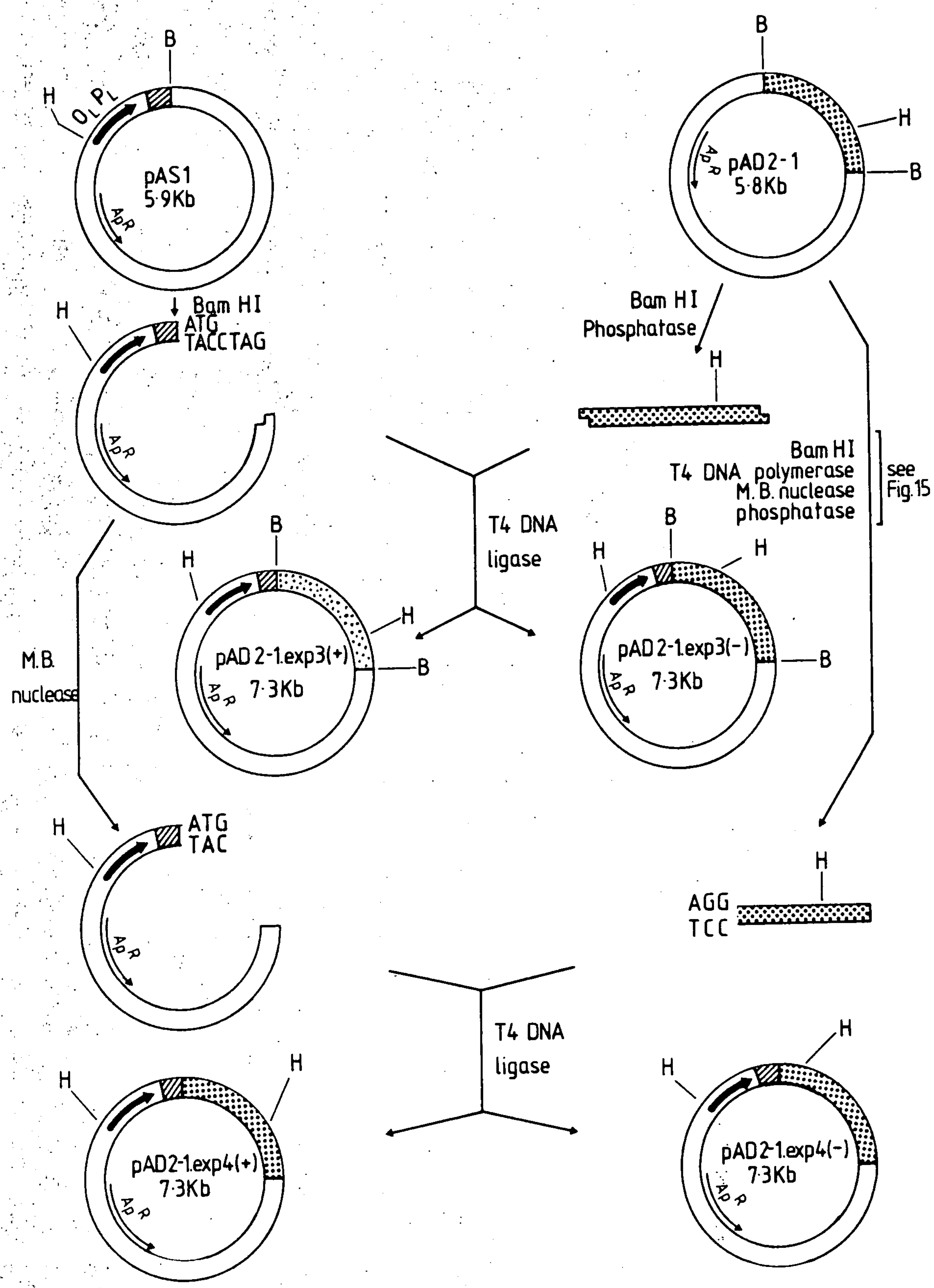


Figure 16. Construction of pAD2-1.exp3 and pAD2-1.exp4 (see text for details).

▨ = ribosome binding site of the phage λ*c*II gene.
Other symbols are used as in Fig. 14. Plasmids pAD2-1.exp3(+) and pAD2-1.exp4(+) contain the cDNA inserts in the correct orientation for expression.

nuclease. pAD2-1 was restricted with BamHI and treated with T4 DNA polymerase in the presence of dTTP. The DNA fragments were blunt-ended with mung-bean nuclease and then treated with alkaline phosphatase. The blunt-ended pAD2-1 insert was ligated in ∿20-fold molar excess to the mung-bean nuclease-treated pAS1. The ligation products were used to transform competent cells of *E.coli* N99λcI$^+$ and *E.coli*. N99λcI857 to ampicillin resistance. 5 out of 40 of the N99λcI857 transformants and 22 out of 198 of the N99λcI$^+$ transformants were positive in a colony hybridisation screen with a ^{32}P-labelled pAD2-1 probe. HindIII digestion of the plasmids from the positive clones, followed by agarose gel electrophoresis, indicated that 2 of the N99λcI857 and 11 of the N99λcI$^+$ plasmids contained the insert in the right orientation for expression. Samples of plasmids obtained from these 11 N99λcI$^+$ clones were transferred into N99λcI857 cells. Since the enzymic reactions used in the construction of these plasmids were more error-prone than those used in the constructions involving straight forward subcloning via cohesive termini, cells harbouring the 13 plasmids with the pAD2-1 insert in the appropriate orientation for vicilin expression were immunologically screened *in situ* for the production of vicilin. Only one plasmid appeared to direct high-level synthesis of a protein which reacted with the antivicilin antibody; it was designated pAD2-1.exp4(+). One of the plasmids with the insert in the opposite orientation was designated pAD2-1.exp4(-).

3.4.5. pAD3-4.exp1.

For the construction of pAD3-4.exp1, a vicilin cDNA encoding a cleavable β : γ endoproteolytic site was inserted into the BamHI site of pAS1. The construction is shown in Fig.17. Since the only vicilin cDNA encoding a cleavable processing site, pAD3-4, lacked some 3'-proximal coding sequence (see Fig.11), an essentially full-length, hybrid cDNA consisting largely of pAD3-4 was constructed as shown in Fig.17 (see Fig.7 for detailed restriction maps of the vicilin cDNAs).

A 1353 bp SalI - BstEII fragment, a 102 bp BstEII-HphI fragment and a 288 bp HphI fragment were isolated on an agarose gel from restriction digests of pAD3-4, pDUB4 and pDUB2 respectively. The 1353 bp and 288 bp fragments were treated with alkaline phosphatase,

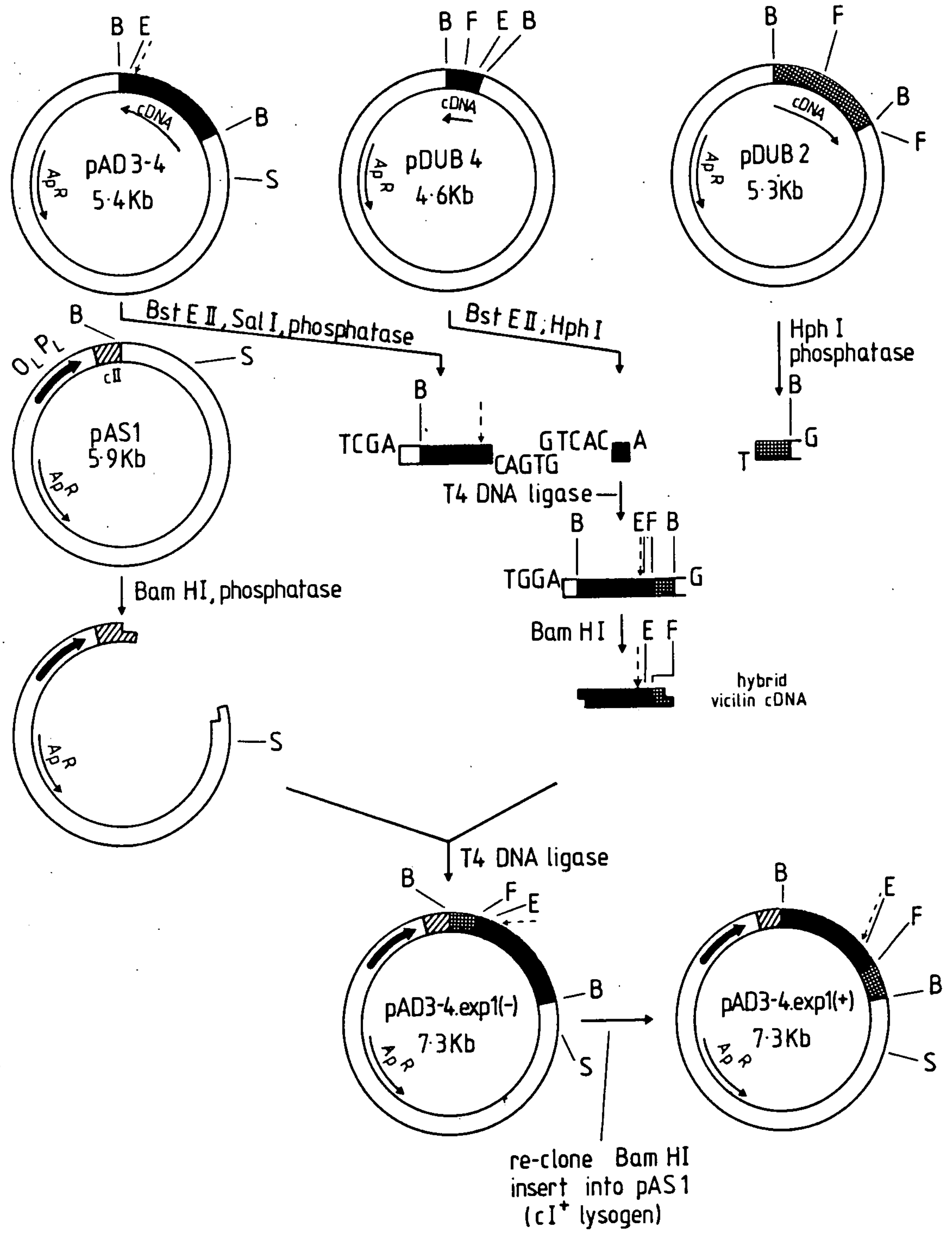


Figure 17. Construction of pAD3-4.expl (see text for details)

■ = vicilin cDNA inserts in pAD3-4 and pDUB4 (overlapping cDNAs from the same gene - see section 3.2.7). The β:γ cleavage site encoded by pAD3-4 is indicated by a dotted arrow.

▦ = vicilin cDNA insert in pDUB2

B = BamHI; E = BstEII; F = HphI; S = SalI.

Other symbols are used as in Fig.16. pAD3-4.expl(+) contains the hybrid cDNA insert in appropriate orientation for expression. The directions of the translational reading frames of the cDNA inserts are indicated by arrows.

ligated in approximately equimolar quantities to the 102 bp fragment, and then cleaved with BamHI. (N.B. Because the HphI cleavage site is removed from the recognition sequence, only one terminus of the 288 bp HphI fragment was compatible with the HphI terminus on the 102 bp BstEII-HphI fragment —— see Fig.17). The resulting tri-hybrid 1447 bp BamHI fragment was purified on an agarose gel and ligated to an estimated 3-fold molar excess of phosphatase-treated, BamHI-linearised pAS1. *E.coli* N99λ*c*I$^+$ was transformed to ampicillin resistance with the ligation products. Out of 1600 colonies screened by colony hybridisation to a ^{32}P-labelled pAD3-4 insert probe, 3 positives were scored. Combined digestion with SalI and XbaI, followed by agarose gel electrophoresis, showed that all 3 plasmids had the insert in the opposite orientation to the direction of transcription. One of the 3 plasmids, designated pAD3-4.expl(-), was selected for further work. To confirm that the hybrid insert had been correctly assembled, cleavage sites for XbaI, BstEII and BglII were mapped on the insert and were found to be consistent with the expected restriction pattern.

The hybrid insert of pAD3-4.expl(-) was excised with BamHI and recloned into a ∿3-fold molar excess of phosphatased, BamHI-linearised pAS1. Competent cells of *E.coli* N99λ*c*I857 and N99λ*c*I$^+$ were transformed to ampicillin resistance with the ligation products. 72 N99λ*c*I857 transformants were obtained and were immunologically screened *in situ* for the production of vicilin. None of the transformants produced any protein which reacted with the antivicilin probe. However, in a parallel colony hybridisation screen with a ^{32}P-labelled pAD3-4 insert probe, 62 positives were scored out of the 72 N99λ*c*I857 transformants. Plasmid DNA from 8 randomly selected transformants which hybridised to the pAD3-4 probe was digested with XbaI and SalI to determine the orientation of their inserts. Agarose gel electrophoresis showed that all 8 plasmids contained the insert in the wrong orientation for expression, i.e. they were identical to pAD3-4.expl(-). Presumably, the remainder of the 62 positive clones also contained the insert in the wrong orientation. Out of 44 N99λ*c*I$^+$ transformants probed with a ^{32}P-labelled pAD3-4 insert probe, 34 positives were scored. 3 out of 8 randomly chosen positive clones were shown to contain plasmids which had the insert in the correct orientation for expression. One of the three plasmids was designated pAD3-4.expl(+).

3.4.6. pAD3-4.exp2(+).

The construction of pAD3-4.exp2(+) involved essentially the replacement of the 5' terminus of the hybrid insert in pAD3-4.exp1(+) by the 5' terminus of the pAD2-1 cDNA, thus removing the vicilin signal peptide sequence present in the pAD3-4 cDNA. The construction is shown schematically in Fig.18. Plasmid pAD3-4.exp1(+) was restricted with XbaI and SalI, and the 1197 bp fragment was purified on an agarose gel. pAD2-1.exp3(+) was similarly restricted with XbaI and SalI. The 6079 bp fragment generated was purified on an agarose gel, phosphatased, and ligated in 3-fold molar excess to the 1197 bp fragment from pAD3-4.exp1(+). The ligation products were used to transform *E.coli*N99λ*c*I$^+$ to ampicillin resistance. Since a directional subcloning procedure was used and the proportion of recombinant plasmids formed was expected to be close to 100%, the usual colony hybridisation screen was omitted. Plasmid DNA from 8 randomly chosen transformants was digested with SalI and XbaI, and analysed by agarose gel electrophoresis which showed that all 8 plasmids contained reconstituted XbaI and SalI sites in the expected positions. One of these plasmids was designated pAD3-4.exp2(+); it effectively contains a 1394 bp tetrahybrid, vicilin cDNA insert comprising fragments originally derived from pAD2-1(476 bp), pAD3-4(453 bp), pDUB4(102 bp) and pDUB2(263 bp).

3.5. Expression of Vicilin Genes in *E.coli.*

3.5.1. Detection of Synthesised Vicilin by *in situ* Colony Immunoassay.

Four different λ*c*I*ts*857 lysogens, transformed with each of the vicilin expression plasmids, were screened *in situ* for vicilin-specific antigenic determinants by reaction with antivicilin IgG as previously described (sections 2.2.24 and 2.2.25). The results (Fig.19) show that all the exp(+) plasmids with the exception of pAD2-1.exp2(+), directed the synthesis of proteins which reacted with the antibody. The relative efficiencies of expression were pAD2-1.exp1(+) >> pAD2-1.exp4(+) ≈pAD3-4.exp1(+) >pAD3-4.exp2(+) >pAD2-1exp3(+). Negative controls (exp(-) plasmids) did not produce any detectable vicilin. The best host strains appeared to be K12ΔH1Δ*trp* and N99λ*c*I857.

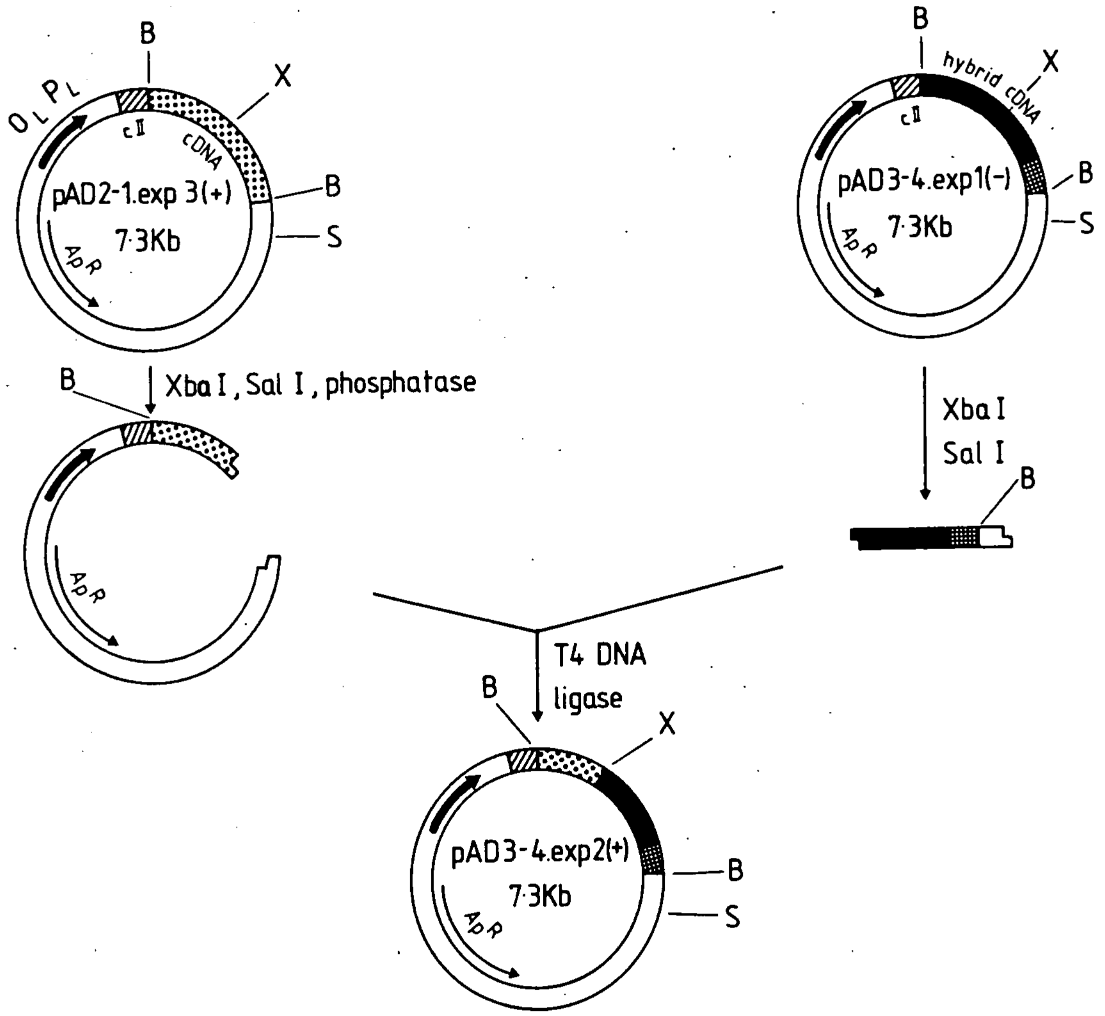


Figure 18. Construction of pAD3-4.exp2(+) (see text for details). Symbols are as used in Figs. 15 and 17. pAD3-4.exp2(+) contains the hybrid vicilin cDNA in the correct orientation for expression.

FIGURE 19. Detection of vicilin synthesis by colony immunoassay. After induction at 42°C for 2.5hr, the colonies were reacted with rabbit, antivicilin IgG.

Lysogenic strains used as hosts:

A and E = K12ΔH1Δ*trp*; B = SG4044{pcI857};
C = N99λ*c*I857; D = N5151 (*c*I857).

Plasmids:

1A - 1D = pAD2-1.expl(+); 1E = pAD2-1.expl(-);
2A - 2D = pAD2-1.exp2(+); 2E = pAD2-1.exp2(-);
3A - 3D = pAD2-1.exp3(+); 3E = pAD2-1.exp3(-);
4A - 4D = pAD2-1.exp4(+); 4E = pAD2-1.exp4(-);
5A - 5D = pAD3-4.expl(+); 5E = pAD3-4.expl(-);
6A - 6D = pAD3-4.exp2(+).

FIGURE 20. Optimisation of conditions for the induction of vicilin synthesis in *E. coli* K12ΔH1Δ*trp*. A : electrophoresis of bacterial cell extracts (equivalent of 400μl of culture per track) on a 12.5% SDS-polyacrylamide gel stained with Kenacid blue. Track 1 contains 30μg of purified vicilin. Extracts of cells transformed with pAD2-1.expl(+) (tracks 2-10) and pAD2-1.expl(-) (tracks 11,12) are shown. Induction procedures were as follows

TRACKS:

2) induced at O.D$_{650}$ = 0.36, switched to 42°C;
3) " , 65 + 42°C (see section 3.5.2.);
4) induced at O.D$_{650}$ = 0.56, switched to 42°C;
5) " , 65 + 42°C;
6) induced at O.D$_{650}$ = 0.72, switched to 42°C;
7) " , 65 + 42°C;
8) induced at O.D$_{650}$ = 0.9 , switched to 42°C;
9) " , 65 + 42°C;
10) uninduced;
11) induced at O.D$_{650}$ = 0.64, switched to 42°C;
12) uninduced.

Prominent bands found specifically in the induced exp(+) tracks are indicated by arrows.

B : Western blot of a duplicate gel reacted with antivicilin IgG.

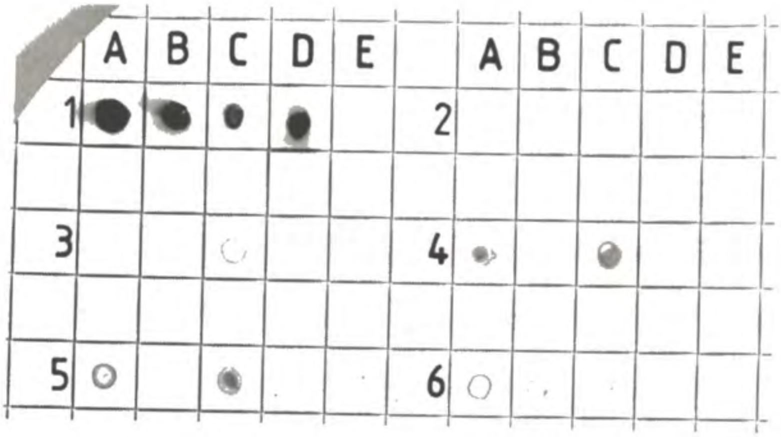

A
B
C
D
E
A
B
C
D
E
1
2
3
4
5
6

Fig. 19

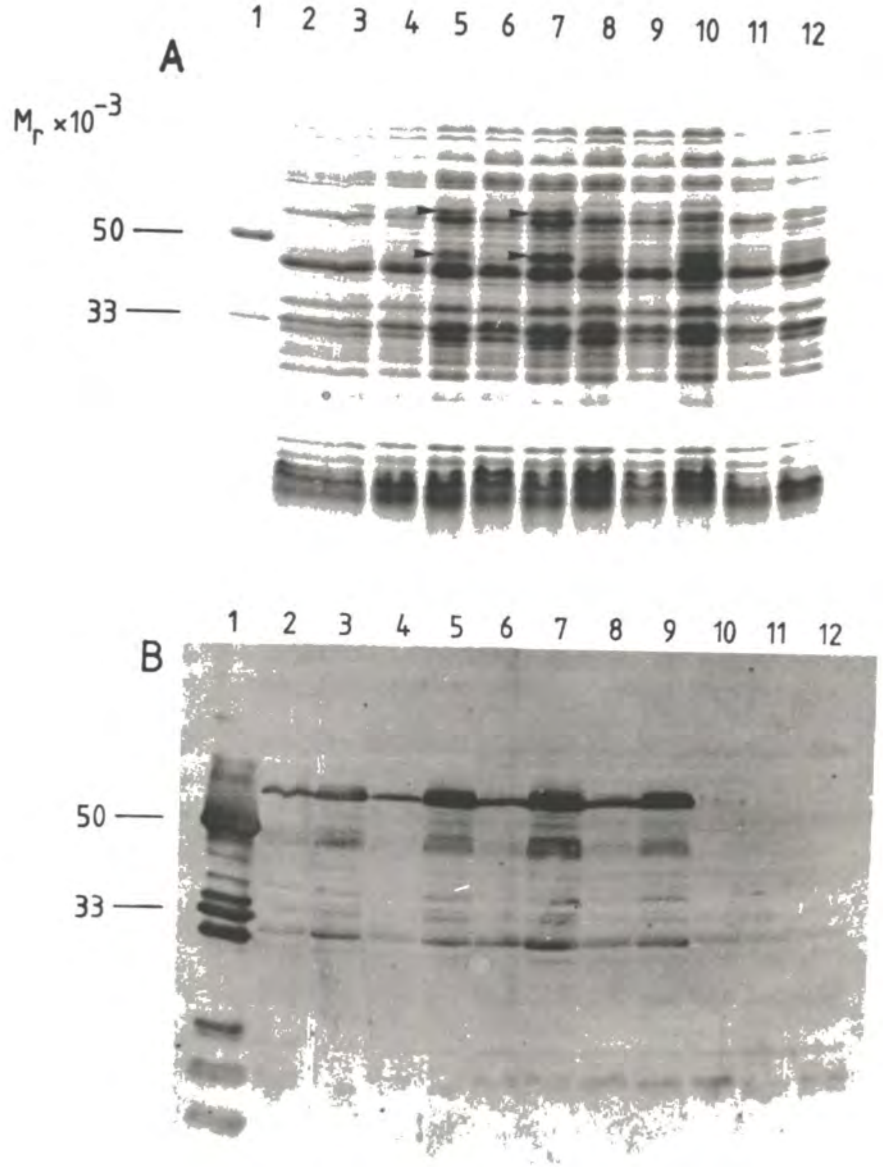

A
1 2 3 4 5 6 7 8 9 10 11 12
M_r ×10^-3
50
33
B
1 2 3 4 5 6 7 8 9 10 11 12
50
33

Fig. 20

3.5.2. Optimisation of Conditions for the Induction of Vicilin Synthesis.

The sizes of the synthesised vicilin molecules were determined by SDS-PAGE analysis of bacterial cell extracts and immunoassay of Western blots. Before comparing the proteins produced by different plasmids, induction conditions were optimised as follows. A culture of K12ΔH1Δ*trp* cells harbouring pAD2-1.expl(+) was grown at 30°C and aliquots were withdrawn for induction at various cell densities as indicated in Fig.20. Induction was effected either by transferring the cultures directly to 42°C or by thorough mixing with equal volumes of L broth prewarmed to 65°C, and then incubating at 42°C (henceforth referred to as the "65+42°C" procedure). After incubation for 2.5hr at 42°C, total cell extracts were subjected to SDS-PAGE analysis. New protein bands of *M*r ∿62000 and ∿48000 were detectable by Kenacid blue staining in the induced cultures but not in uninduced cells or in cells transformed with pAD2-1.expl(-)(Fig.20A). These bands were most prominent in cells induced at O.D.$_{650}$'s of 0.56 and 0.72 by the "65+42°C" procedure. Fig.20B shows a Western blot of a duplicate gel screened with anti-vicilin IgG. The antibody reacted strongly with protein bands of ∿62000-*M*r present specifically in the induced cells harbouring pAD2-1.expl(+); there were also weaker reactions with a number of lower *M*r bands. The strongest reaction was with proteins from cells induced at an O.D.$_{650}$ of 0.72 by the "65+42°C" method. These conditions were adopted for all subsequent inductions.

3.5.3. Comparisons of Vicilin Synthesis Directed by Different Plasmids.

E.coli N99λ*c*I857 cells harbouring the various vicilin expression plasmids were induced as described in the preceding section, and K12ΔH1Δ*trp* cells harbouring pAD2-1.expl(+) and pAD2-1.exp4(+) were similarly induced for comparison of the effects of the host strain on vicilin synthesis. The bacterial pellets obtained from the N99λ*c*I857 cultures were noticeably smaller than the K12ΔH1Δ*trp* pellets, and this was reflected in an analysis of the total cell extracts by SDS-PAGE (Fig.21A). The gel shows that considerably less protein was recovered in general from the induced N99λ*c*I857 cultures compared to an uninduced culture or to the induced K12ΔH1Δ*trp* cultures. The only protein band detected specifically

FIGURE 21. Comparisons of vicilin synthesis by different expression plasmids. A : electrophoresis of bacterial cell extracts on a 15% SDS-polyacrylamide gel stained with Kenacid blue. Track 1 contains 20μg of purified vicilin. Extracts of bacterial cells (equivalent of 400μl of culture per track) transformed with the following plasmids are compared.

Track:

2) pAD2-1.exp1(+) in K12ΔH1Δ*trp* , induced;

3) pAD2-1.exp1(+) in N99λ*c*I857 , induced;

4) pAD2-1.exp2(+) in N99λ*c*I857 , induced;

5) pAD2-1.exp3(+) in N99λ*c*I857 , induced;

6) pAD2-1.exp4(+) in K12ΔH1Δ*trp* , induced;

7) pAD2-1.exp4(+) in N99λ*c*I857 , induced;

8) pAD3-4.exp1(+) in N99λ*c*I857 , induced;

9) pAD3-4.exp2(+) in N99λ*c*I857 , induced;

10) pAD2-1.exp1(+) in N99λ*c*I857 , uninduced;

11) pAD2-1.exp1(-) in K12ΔH1Δ*trp* , induced.

A ∿62000-*M*r vicilin fusion protein is indicated by an arrow.

B : Western blot of a duplicage gel reacted with activicilin IgG.

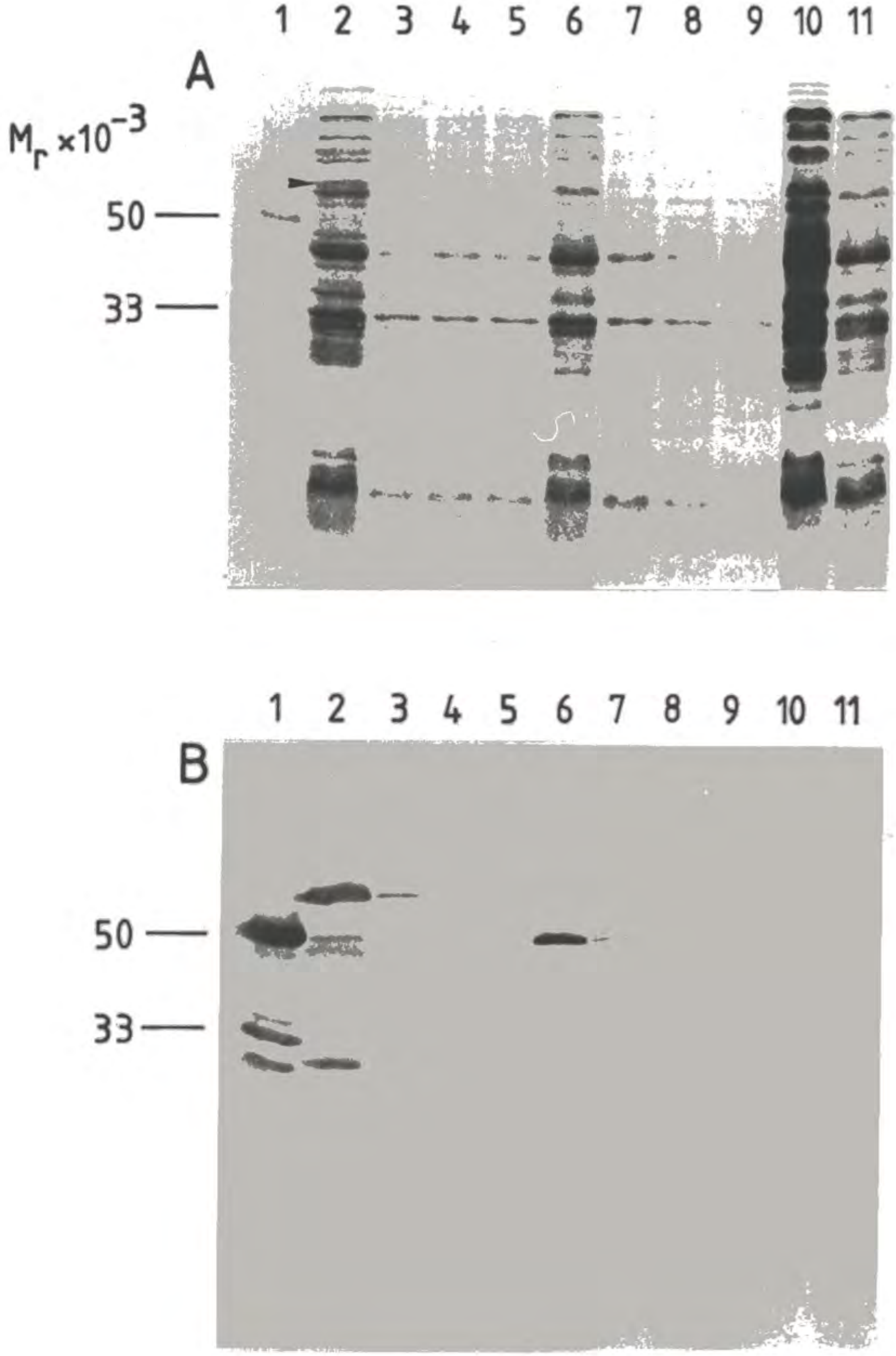


Fig. 21

in the induced cultures was the ∿62000-*Mr* protein in cells harbouring pAD2-1.expl(+). A duplicate gel was subjected to Western blotting and screened with antivicilin IgG (Fig.21B). The antibody reacted with proteins in induced cells transformed with pAD2-1.expl(+), pAD2-1.exp3(+), pAD2-1.exp4(+) and pAD3-4.expl(+). The tested K12ΔH1Δ*trp* host cells accumulated more (3-10-fold) vicilin than the N99λ*c*I857 cells transformed with the same plasmids.

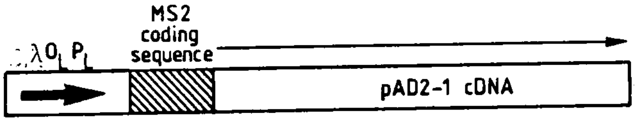
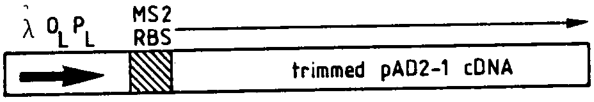
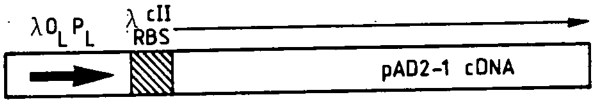
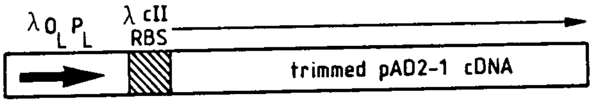
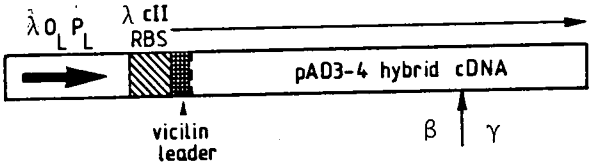
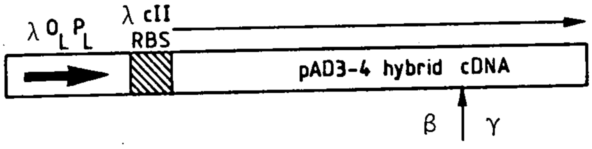
No vicilin was detectable in induced cells harbouring pAD2-1.exp2(+), pAD3-4.exp2(+) and pAD2-1.expl(-), or in uninduced cells harbouring pAD2-1.expl(+). The relative efficiencies of vicilin synthesis directed by the various plasmids were similar to that observed in the colony immunassay (section 3.5.1.) and is summarised in Table 8 together with the *Mr*'s of the synthesised vicilins.

The above experiment was repeated using K12ΔH1Δ*trp* cells as hosts for the expression plasmids. The gel showed the expected presence of the ∿62000-*Mr* protein synthesised by pAD2-1.expl(+) but no other unique bands were detectable in the induced cells containing the other expression plasmids (Fig.22A). The Western blot showed that all the induced exp(+) plasmids except pAD2-1.exp2(+) directed the synthesis of proteins which reacted with the antivicilin IgG (Fig. 22B). The yields of bacteria-synthesised vicilin were estimated by comparison with known amounts of vicilin samples and are indicated in Table 8.

3.5.4. Effects of Temperature on the Growth of two *E.coli* Lysogens.

Fig. 23 shows the growth rates of K12ΔH1Δ*trp* and N99λ*c*I857 cells, both harbouring pAD2-1.exp4(+), when incubated at 30°C and upon induction by the "65+42°C" method (see Section 3.5.2.). Both cultures displayed typically sigmoidal growth curves at 30°C as did the K12ΔH1Δ*trp* culture at 42°C, reaching plateaus at O.D.$_{650}$'s of ∿1.1 and 0.85 respectively. By contrast, the cell density of the N99λ*c*I857 culture decreased sharply when incubated at 42°C. After dropping to an O.D.$_{650}$ of ∿0.11, exponential growth was resumed reaching a plateau of O.D.$_{650}$ ≃0.22.

Table 8. Summary of constructions and properties of various expression plasmids.

Plasmid	Construction	Approx. *Mr* of product	Approx. R.E.(a)
pAD2-1.expl(+)		62000	1.0
pAD2-1.exp2(+)		-	0
pAD2-1.exp3(+)		50000	0.2
pAD2-1.exp4(+)		50000	0.3
pAD3-4.expl(+)	 λO$_L$ P$_L$ λ cII RBS pAD3-4 hybrid cDNA vicilin leader β γ	47000	0.2
pAD3-4.exp2(+)		47000	0.2

a. R.E. = relative efficiency of vicilin synthesis in the bacteria. ∿25μg of vicilin per ml of culture were synthesised by pAD2-1.expl(+) as estimated by comparisons with standard amounts of purified vicilin on a Western blot (see Fig. 22B). "Trimmed" cDNA refers to the removal of linker and signal peptide sequences from the pAD2-1 cDNA inserts so that the N-terminal codon of mature vicilin was positioned at the 5' terminus of the cDNA. Horizontal arrows over the vicilin inserts indicate the direction of the cDNA translational reading frame. exp(-) constructions had the cDNAs in the opposite orientation. Vertical arrows in the pAD3-4.expl(+) and pAD3-4.exp2(+) constructs indicate cleavable sites encoded by the cDNAs.

FIGURE 22. Comparisons of vicilin synthesis by different expression plasmids maintained in strain K12ΔH1Δ*trp*. A: electrophoresis of bacterial cell extracts on a 12.5% SDS-polyacrylamide gel stained with Kenacid blue. Tracks 1, 2 and 3 contain 5, 10 and 20 μg of purified vicilin respectively. Extracts of cells (equivalent of 400μl of culture per track) transformed with the following plasmids are compared.

Track:

4) pAD2-1.exp1(+), induced;

5) pAD2-1.exp2(+), induced;

6) pAD2-1.exp3(+), induced;

7) pAD2-1.exp4(+), induced;

8) pAD3-4.exp1(+), induced;

9) pAD3-4.exp2(+), induced;

10) pAD2-1.exp1(-), induced;

11) pAD2-1,exp1(+), uninduced.

A ∿62000-*M*r vicilin fusion protein is indicated by an arrow.

B:Western blot of a duplicate gel reacted with antivicilin IgG.

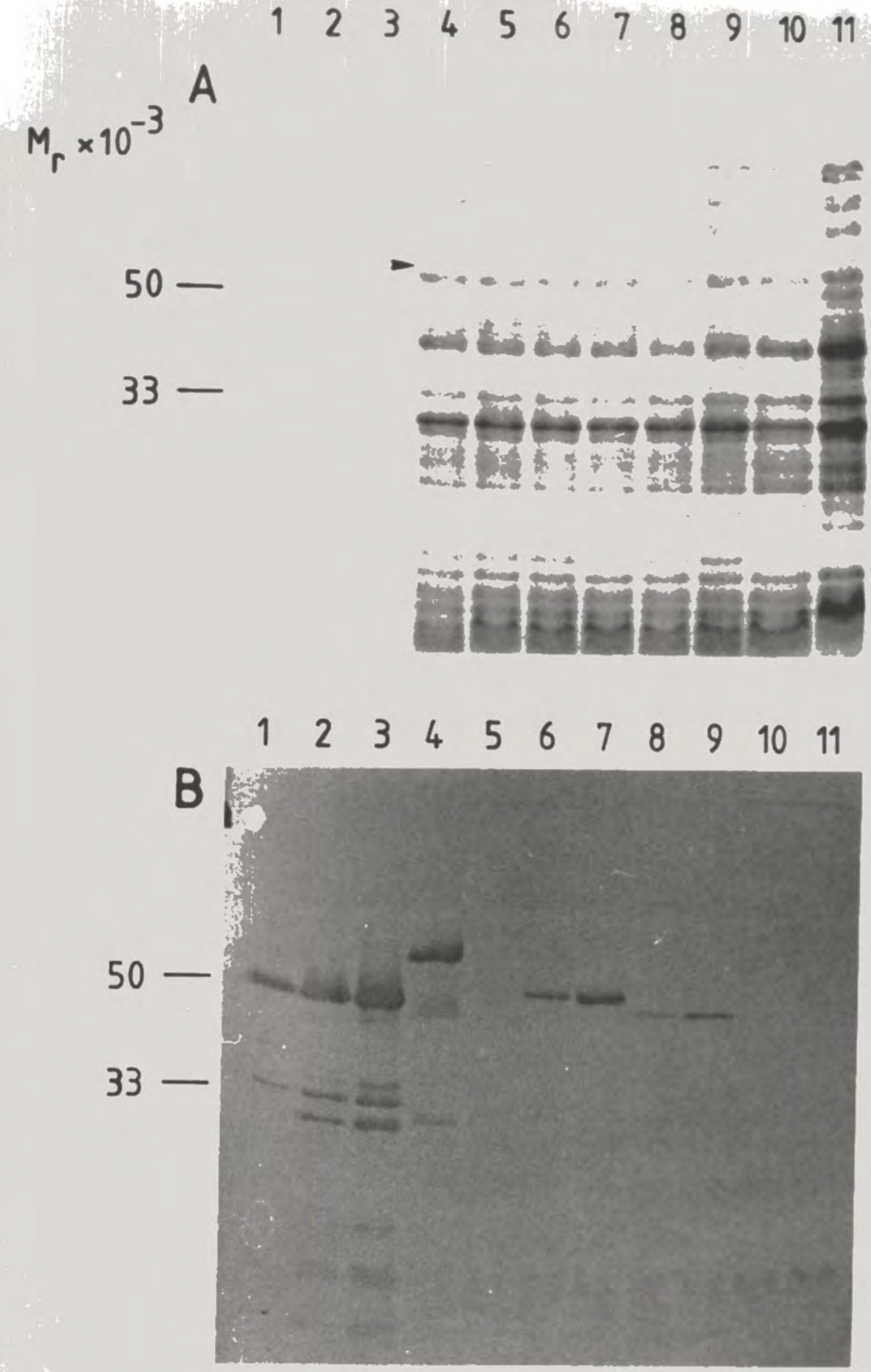


Fig. 22

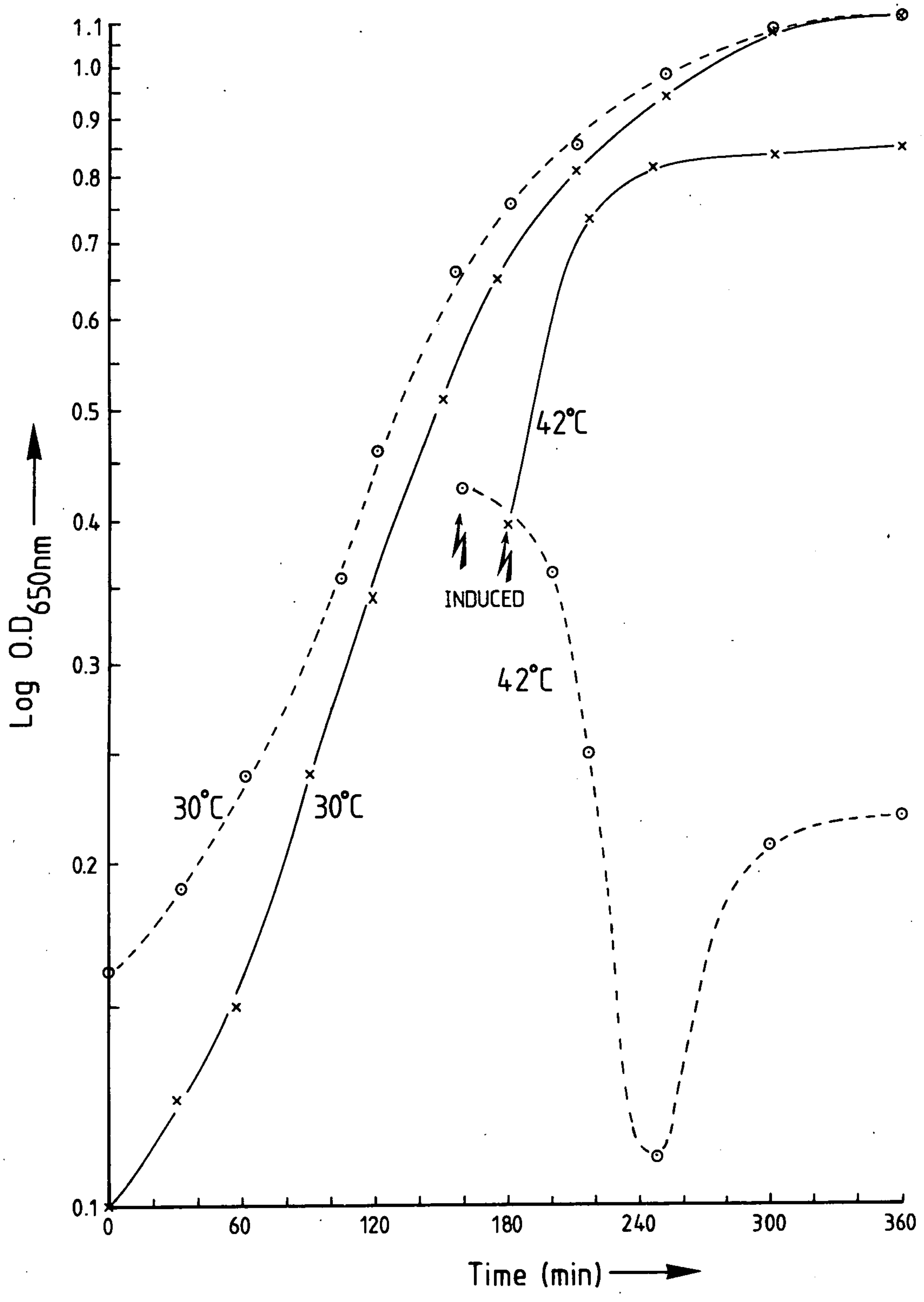


FIGURE 23. Effect of temperature on the growth of two *E.coli* lysogens. Cultures were grown initially at 30°C; then, at O.D$_{650}$ ≃0.66, aliquots were withdrawn and induced by the "65 + 42°C" procedure (see section 3.5.2.). The O.D dropped immediately by ∿40% upon induction.

ʘ----ʘ = O.D$_{650}$ of strain N99λ*c*I857 transformed with pAD2-1.exp4(+).

x——x = O.D$_{650}$ of strain K12ΔH1Δ*trp* transformed with pAD2-1.exp4(+)

3.6. Reconstruction of pAD2-1.exp2(+).

The failure of pAD2-1.exp2(+) to synthesize detectable amounts of vicilin prompted the reconstruction of the plasmid. In this reconstruction, the vector-5'cDNA junction in the original construction was subcloned into the appropriate segment of pAD2-1.exp1(+), thus effectively replacing the $\lambda O_L P_L$ region in the original plasmid with the homologous region from pAD2-1.exp1(+). The reconstruction is shown schematically in Fig.24. K12ΔH1Δ*trp* cells transformed with the reconstructed plasmid did not produce any vicilin as judged by SDS-PAGE analysis and immunoassay of Western blots.

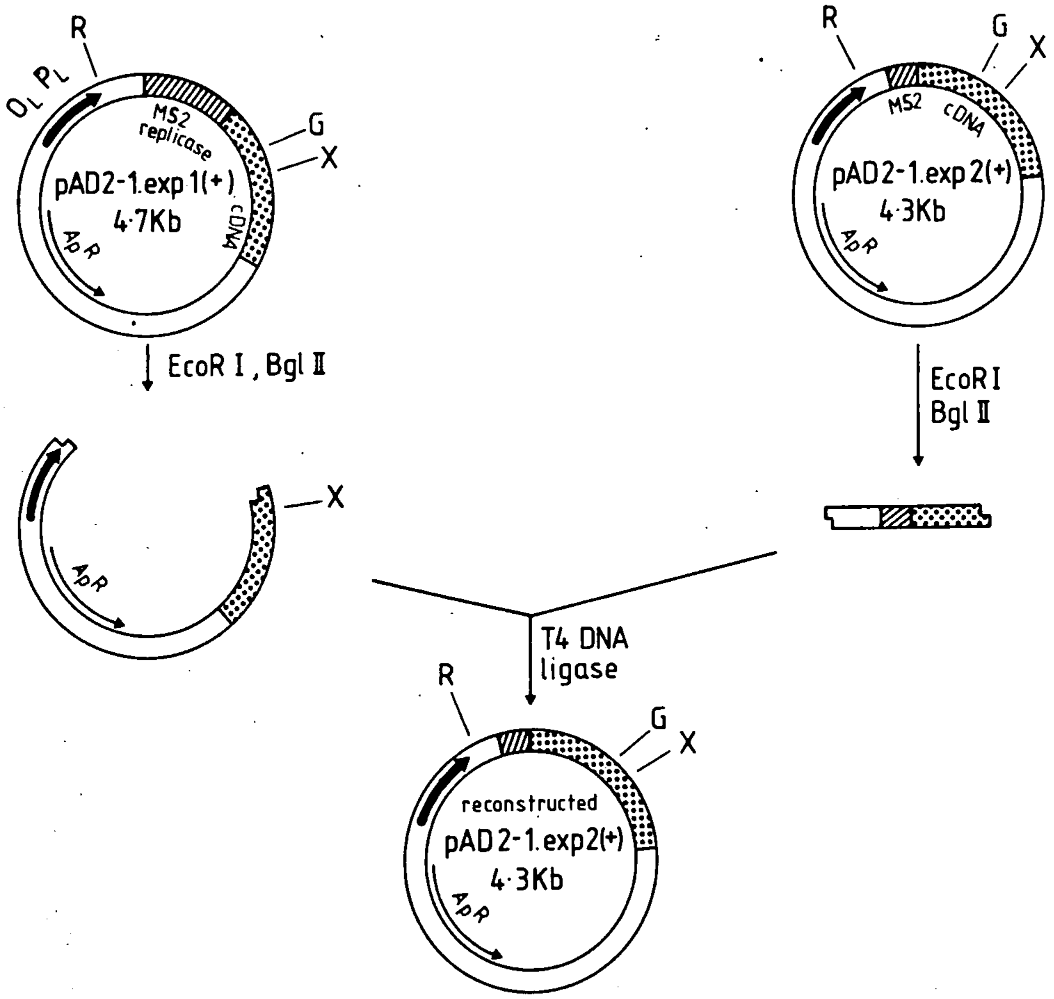


Figure 24. Reconstruction of pAD2-1.exp2(+) (see text for details). Other symbols are as used in Figs. 14 and 15. The plasmid construction was verified by restriction with EcoRI and XbaI, followed by agarose gel electrophoresis.

4. DISCUSSION.

4.1. General Assessment of Methods used for the Construction of the cDNA Library.

Several methods have been developed for the construction of cDNA libraries (for reviews see Williams, 1981; Maniatis *et al.*, 1982; Forde, 1983a, b). In the present work, a cDNA library was constructed from size-fractionated cDNAs, synthesised from pea cotyledon mRNAs, and cloned into the BamHI site of pBR322 using BamHI linkers. The initial screening of the library was based on the inactivation of the tetracycline resistance (Tc^R) gene in the vector. An unsatisfactory feature of the results obtained was the small proportion of tetracycline-sensitive (Tc^S) clones scored among the ampicillin-resistant (Ap^R) transformants (see Table 1). Since the high percentage of $Ap^R.Tc^R$ colonies obtained must have arisen from cells which had been transformed with either recircularised or oligomeric plasmid molecules, the strategies adopted for minimizing the formation of these plasmid species need to be re-examined.

One strategy involved phosphatase treatment of the vector to prevent self-ligation. Although agarose gel electrophoresis of the ligation products formed between the phosphatase-treated pBR322 and the cDNA indicated the absence of recircularised or oligomeric vector molecules, these molecular species were probably present in quantities too low to be visualised on the gel but in sufficient amounts to yield a high background of nonrecombinant transformants. Incomplete linearisation in the initial restriction of the pBR322 DNA may also have contributed to the presence of trace amounts of nonrecombinant plasmid molecules. A number of trial ligation and transformation experiments (Goodman and McDonald, 1979; Maniatis *et al.*, 1982) may be performed to ascertain that the linearised vector is not contaminated with uncut DNA and to monitor the efficiency of the phosphatase treatment. The difficulty experienced in this work in screening for recombinant clones would support the suggestion that such trial experiments should be made an integral part of the cDNA cloning protocol.

The alternative strategy employed for reducing the production of nonrecombinant clones involved the recovery of chimaeric plasmid molecules from an agarose gel prior to transformation. Though this procedure should theoretically have excluded any recircularised or oligomeric plasmids, it proved to be even less successful than the phosphatase treatment procedure. The presence of contaminating nonrecombinant molecules in the DNA used for transformation may be explained, at least partially, by the inability of agarose gel electrophoresis to effect an absolute separation of DNA molecules. This contention is supported by the fact that when cDNA inserts, isolated from agarose gels, are labelled to high specific activity and used to probe Southern blots, some hybridisation to vector fragments present on the blots is invariably observed (this is discussed in more detail in section 4.3). It is also possible that in an effort to maximize the yield of hybrid plasmids, some contamination by nonrecombinant plasmid molecules occurred in the process of recovering the DNA from the gel.

An important factor which must have contributed to the high background of nonrecombinant plasmids was recognised only after completion of the cloning experiment. Using the same procedure described in section 2.2.10. for cDNA synthesis, I.M. Evans (pers. commun.) has consistently found that the yields of double-stranded cDNA (ds-cDNA) were 20-30% of the starting mass of poly$(A)^+$ mRNA. Percentage yields of similar magnitude are also reported by Maniatis *et al.*, (1982). Thus, the maximum yields of ds-cDNA synthesised from 6μg of poly$(A)^+$ RNA would be ∿1.5μg which would be reduced to <1μg after the recovery of size-fractioned molecules from an agarose gel as carried out in this work. Clearly therefore, the calculated recoveries of the 1-2Kb (∿5μg) and >2Kb(∿2μg) cDNA size classes were over-estimated. Consequently, in the ligation of the cDNA to the plasmid vector, the vector must have been present in ∿10x higher excess than was estimated, i.e. in 60-90-fold molar excess over the cDNA. The use of such a large excess of the vector would only have served to increase the proportion of nonrecombinant molecules formed. Bearing in mind that this miscalculation would have had a greater effect on the strategy involving phosphatase treatment, and yet that strategy proved to be more efficient than the isolation of chimaeric plasmids from a gel, it may be concluded that the phosphatase treatment procedure is more effective in minimizing the production of nonrecombinant transformants in cDNA cloning.

An additional problem encountered in screening the library was that only about half of the Tc^S transformants hybridised to a cotyledon poly(A)$^+$ RNA probe. The origins of the plasmids which did not hybridise to the probe was not investigated in depth, but the possibility that they were transcribed from contaminating rRNA molecules in the poly(A)$^+$ RNA preparation can be discounted since none of the recombinant plasmids hybridised to a cotyledon rRNA probe. It is possible that they arose from oligomeric plasmid forms which were subject to various molecular rearrangements or deletions effected either by enzymic reactions *in vitro* or inside the bacterial cells.

Out of 59 mRNA-hybridising clones obtained from one cDNA library, 39 (68%) were shown by colony hybridisations to contain legumin or vicilin cDNAs (Figs.4B and 4D), reflecting the fact that the mRNA population of the developing seed is highly enriched for storage protein messages (Morton *et al.*,1983). Restriction mapping and DNA sequencing revealed that no full-length cDNA molecules had been cloned and that, in addition, sequence artefacts had been incorporated in some of the clones.

The failure to obtain full-length cDNAs was, in fact, partly a feature of the cloning procedure used. This procedure relies for second strand cDNA synthesis on priming by the hairpin loop structures at the 3' end of the first cDNA strands. Digestion of the hairpin loops with S1 nuclease (which must precede the insertion of the cDNA into the vector) invariably removes portions of the cDNA corresponding to the extreme 5' end of the mRNA. The use of S1 nuclease may effect further losses of cloned DNA sequences due to the tendency of the enzyme to "nibble" at the termini of ds-DNA molecules (Shenk *et al.*,1975) and, when used in high concentrations, to cleave transiently single-stranded regions caused by partial denaturation of ds-DNA (Lathe *et al.*,1983). It is possible that over-digestion with S1 nuclease may have caused the loss of 3' sequences from several of the cDNAs; alternatively, this may also have resulted from incomplete second strand synthesis followed by legitimate single-stranded scission of the protruding ends (Forde, 1983a).

Alternative methods for ds-cDNA synthesis which obviate the requirement for S1 nuclease digestion have been developed. In these methods, the 3' end of the first strand is tailed with dT (Rougeon *et al.*,1975) or dC residues (Land *et al.*,1981), and the second strand is then synthesised using an oligo(dA) or oligo(dG) primer respectively. A second set of homopolymer tracts is attached to the resulting duplex DNA which is then annealed to a plasmid vector tailed with complementary nucleotide residues. The main attraction of these procedures and variations designed to improve the cloning efficiencies (Okayama and Berg, 1982; Heidecker and Messing, 1983) is that they enable the cloning of full-length cDNAs.

However, homopolymer tailing cloning methods suffer certain disadvantages compared to the methods employing restriction enzyme linkers : (i) the use of linkers enables the cDNA insert to be precisely excised and hence easily purified which is not always possible with the tailing methods. Even when the particular tailing strategy used reconstitutes restriction sites at each end of the inserted DNA thus allowing resection of the cDNA (e.g. Villa-Komaroff *et al.*,1978), the insert is excised with the homopolymer tails still attached. This may present problems if the labelled cDNA is used in hybridisation experiments in which the target DNA contains nucleotide tracts complementary to the cDNA tails. Another problem, of which at least one example is known, is that the homopolymer tails may prove refractory to enzymatic removal (Edens *et al.*, 1982). (ii) Under the conditions normally used for the terminal transferase reactions, homopolymer tails may be added at internal nicks in the ds-cDNA (Nelson and Brutlag, 1979) and may thus cause serious losses of cDNA sequences (e.g. Kupper *et al.*,1981). By contrast, cDNA clones obtained by the linker method should theoretically contain inserts as long as the starting ds-cDNA material (Williams, 1981). (iii) the infectivity of annealed recombinant plasmids is considerably lower than that of covalently closed ones, and consequently, the linker methods require a much smaller amount of mRNA to generate a given number of cDNA clones (Williams, 1981).

In view of the above considerations, it would seem desirable to develop a method for cloning full-length cDNAs which take advantage of the relative merits of the linker methods. The following cloning scheme is therefore proposed : (i) synthesize the first cDNA strand on a poly(A)$^+$ mRNA template by standard procedures (Maniatis *et al.*,1982). (ii) Attach linkers to the cDNA-mRNA hybrid using a combination of T4 DNA ligase and T4 RNA ligase. (iii) hydrolyse the mRNA with alkali and synthesize the second cDNA strand using *E.coli* DNA polymerase 1. The linkers attached to the 3' end of the first cDNA strand should serve as a suitable primer for second strand synthesis. Alternatively, replace the mRNA strand using the combined activities of RNase H and *E.coli* DNA polymerase 1 (as in Okayama and Berg, 1982). (iv) Digest the linkered ds-cDNA with the appropriate restriction enzyme to generate cohesive termini and ligate into a phosphatase-treated plasmid vector.

The only questionable step in the outlined procedure concerns the efficiency of the ligation of linkers to the cDNA-mRNA hybrid. However, the properties of T4 DNA ligase (reviewed by Engler and Richardson, 1982) suggest that the enzyme should catalyse the blunt-end ligation of the linkers to the DNA strand of the hybrid. T4 RNA ligase is included in the ligation reaction since it is known to stimulate the activity of T4 DNA ligase on blunt-ended DNA molecules approximately 20-fold (Sugino *et al.*,1977a), and is itself active with both DNA and RNA substrates (Sugino *et al.*,1977b; Brennan *et al.*,1983; Harrison and Zimmerman, 1984). The sequence artefacts evident in some of the characterised clones will be discussed later in the context of specific examples,but it may be noted here that the incidence of some of these artefacts may be reduced by the use of methods which do not rely on self-priming for second strand cDNA synthesis.

4.2. Analysis of Legumin cDNAs.

Legumin cDNAs in the clone bank were identified by hybridisation to the cDNA insert from pDUB3 (formerly pRC2.11.7.— Croy *et al.*,1982) and the sequences of two overlapping clones, pAD4-4 and pAD10-5 were determined. The composite sequence of pAD4-4 and pAD10-5 extends over almost 90% of the coding sequence of a legumin α-β subunit pair as well as containing the 3' untranslated region of the message. Aligning the cDNA sequences with the complete sequence of a legumin genomic clone (Lycett *et al.*, 1984a) shows that the 5' terminus of pAD10-5 coincides with the codon for the forty-seventh amino acid residue of mature legumin.

pAD4-4 and pAD10-5 are completely homologous in the region where they overlap (550 bp) with the exception of two nucleotide substitutions (see Fig.10), and it was initially uncertain whether these mismatches reflected real differences between two legumin genes or were due to errors by reverse transcriptase during cDNA synthesis (see Lycett *et al.*,1984b). However, comparisons with the sequence for a complete legumin (*leg A*) gene (Lycett *et al.*,1984a) reveal that the pAD10-5 sequence is completely homologous to the sequence of the *leg A* gene and is probably, therefore, derived from it, whereas the pAD4-4 sequence differs in five positions from the *leg A* sequence and is likely to have originated from a different gene. Comparisons of the pDUB3/pDUB1 composite sequence (Croy *et al.*,1982) with the pAD4-4 and pAD10-5 sequences (Fig.10) indicate that pDUB3 and pDUB1 were derived from yet another gene. The 3' untranslated region of pAD4-4 is identical to that of pDUB3/1 (N.B. this region is missing from the pAD10-5 cDNA). As described previously (Lycett *et al.*,1983b), this region contains multiple and overlapping polyadenylation signal sequences, but on the evidence of the cDNA sequences obtained to date, these do not give rise to variability in the polyadenylation of legumin messages.

A striking difference between the pAD4-4 and pAD10-5 sequences on the one hand, and the pDUB3 sequence on the other is the presence in the former of three 54 bp direct repeats in the region coding for the legumin α-subunit whereas only half of one repeat sequence unit is present in the pDUB3 cDNA (see Fig.10). These repeats are imperfect : the second repeat unit contains an internal hexanucleotide duplication and thus, in fact, comprises 60 bp; moreover, apart from this internal duplication, the first and second repeats are more similar to each other (∿96% homology) than to the third (∿70% homology). This suggests that the duplication event which gave rise to the first and second repeats occurred after an initial duplication which gave rise to the third repeat.

Fig.25 compares the amino acid sequences deduced from the legumin cDNAs with the partial protein sequences determined analytically from purified legumin subunits (data taken from Lycett *et al.*,

```
cDNA                                                             ALSR
LEG α    ▌LREQP QQNEC QLERL NALQP DNRIE SEGGF IETWN PNNNE FRECG LD.L
                                                       KQ

cDNA      ATLQR NALRR PYYSN APQEI FIQQG NGYFG MVFPG CPETF EEPQE SEQGE
LEG α                   ... ....               .... ..... ..... .....

cDNA      GRRYR DRHQK VNRFR EGDII AVPTG IVFWM YNDQD TPVIA VSLTD IRSSN
LEG α         . ..... .  .  ..... ...         ..... .

cDNA      NQLDQ MPRRF YLAGN HEQEF LQYQH QQGGK QEQEN EGNNI FSGFK RDFLE
LEG α      .... ..... ..... .   . ..... ...          .... ..... .S...
               -  APG              R F                    L

cDNA      DAFNV NRHIV DRLQG RNEDE EKGAI VKVKG GLRII SPPEK QARHQ RGSRQ
                                                S                 K
LEG α     ..... ..... ..... ..... .         . ..... .....
                                               K

cDNA      EEDED EEKQP RHQRG SRQEE EEDED EERQP RHQRR RGEEE EEDKK ERRGS
                                                                  G
LEG α                                      ... ....               ....

cDNA      QKGKS RRQGD NGLEE TVCTA KLRLN IGPSS SPDIY NPEAG RIKTV TSLDL
LEG α     ..... ..... .
LEG β                 ▌.... ..... ..... ..... ...L. ..... .L... .....

cDNA      PVLRW LKLSA EHGSL HKNAM FVPHY NLNAN SIIYA LKGRA RLQVV NCNGN
                              T
LEG β     ..... ..... ..... ..... ..... .....        ...   .... .....
                             R

cDNA      TVFAG KLEAG RALTV PQNYA VAAKS LNDRF SYVAF KTNDR AGIAR IAGTS
                                     N
LEG β              .. ..... ..... ...                .... ..... L....

cDNA      SVINN LPLDV VAATF KLQRN EARQL KSNNP FKFLV PARQS ENRAS A
             D              N  D
LEG β     ..... ..... ...      . .....         ... ....     .. V
```

Figure 25. Comparisons of protein sequences predicted from the legumin cDNAs (see Fig.10) with partial amino acid sequences determined directly from purified α- and β- subunits. The direct amino acid sequence data are taken from Lycett *et al.* (1984b), and include 9 residues (underlined) from the sequence of glycinin. Symbols are as used in Fig.10 and blank spaces in the analytically determined protein sequences indicate unsequenced regions. ▌ indicates that the N-terminus was determined directly.

1984b). There is almost complete homology between the predicted and determined sequences which confirms that the cDNAs code for precursors of the legumin subunits. Furthermore, it can be seen that the tandem repeats in pAD4-4 and pAD10-5 do not disrupt the translational reading frame of the cDNA sequences, and thus give rise to three amino acid repeats in the predicted protein sequence. Each protein repeat is extremely hydrophilic, containing a high proportion of basic residues over the first half of the repeat followed by a preponderance of acidic residues. Though the relevant region in the legumin acidic subunit has not been fully sequenced, the available amino acid sequence data (Lycett *et al.*, 1984b; Fig.25) confirms the presence of part of these repeats.

A fundamental question arises as to whether the absence of the repeats from the pDUB3 cDNA reflects genuine differences between two types of legumin genes or whether it is artefactual. It is known that several copies of the legumin genes exist in the pea genome — estimates range from three to four (Croy *et al.*, 1982) to at least seven (Shirsat, 1984). Furthermore, Matta *et al.* (1981) have demonstrated that legumin acidic subunits show substantial *M*r variation and are considerably more heterogeneous than the basic subunits. The decrease in *M*r of ∿6000 resulting from the deletion of the repeats in pDUB3 closely matches *M*r differences observed between the 35000 - 38500-*M*r and 43000-*M*r acidic legumin subunits (Matta *et al.*,1981).

The extreme hydrophilicity of the repeats suggests a location on the surface of the folded protein, and it is conceivable that significant variation in the surface morphology of different subunits could be tolerated without disrupting the core structure of these proteins. Moreover, there is evidence that this region in the legumin molecule is subject to variation since the corresponding sequence in soybean legumin (Nielson,1984) comprises a single copy of a sequence showing partial homology to the pea legumin repeats. The above considerations originally led us (Lycett *et al.*,1984b) to favour the view that the two types of legumin cDNAs represented transcripts from different types of legumin genes, although the possibility that pDUB3 represented an artefact could not be discounted.

To try and resolve this ambiguity, replica restriction digests of pea, genomic DNA were probed with cDNA fragments corresponding respectively to regions encoding the legumin basic subunit, the repeats and the acidic subunit region upstream from the repeats. The rationale for this experiment was that legumin gene fragments which lacked the repeats ought to be identifiable by hybridisation to the acidic and/or basic subunit probes and concommitant absence of hybridisation to the probe for the repeats. The results showed that all the bands which hybridised prominently to the acidic and basic subunit probes also hybridised strongly to the repeats probe (see Fig.13). This suggests that all the major legumin genes detectable by Southern blot hybridisation to the legumin cDNAs do contain the repeats. It is, nevertheless, possible that certain legumin genes were undetected by the cDNA probes used, and these may, or may not, contain the repeats. Indeed, a 1.5 Kb HindIII fragment (not detectable under the conditions used by Croy *et al.*, 1982) hybridised to the probes for the acidic and basic subunits at an intensity of ∿0.75 gene equivalents (assuming that each of the four EcoRI fragments correspond to ∿1 gene equivalent —— Croy *et al.*, 1982), and hybridised even more weakly to the probe for the repeats. This result suggests that the 1.5 Kb fragment is less homologous to the cDNAs than the gene fragments previously identified by Croy *et al.* (1982), and furthermore, that that particular fragment may lack a part of the repeats. In fact, preliminary data from the sequencing of different legumin genomic clones confirm that the 1.5 Kb HindIII fragment carries a legumin pseudogene (present on the ∿13 Kb EcoRI fragment in addition to a normal legumin gene) which shows significant sequence divergence from the other legumin genes and contains only part of the repeats (R.Croy, pers. commun.). However, the overall pDUB3 cDNA sequence is highly homologous to that of the other legumin cDNAs, and the chromosomal gene from which it was transcribed would be expected to hybridise strongly with the pAD4-4 and pAD10-5 cDNA probes used here.

Even if, as is suggested by the present data, all the legumin chromosomal genes contain the repeats, it may still be argued that the pDUB3 cDNA accurately represented the structure of an mRNA molecule from which the repeats had been spliced out. The occurrence of alternative splicing of primary transcripts to generate different gene products has been documented in animal and viral gene expression (Darnell, 1982 and refs. therein), and Craik *et al.*, (1983) have postulated the "sliding" of intron-exon junctions as a mechanism for generating sequence polymorphisms in protein families. However, the occurrence of similar processes in the legumin genes is unlikely since an examination of the sequences in pAD4-4 and 10-5 corresponding to the sequence missing from the pDUB3 cDNA (Fig.10) does not reveal the characteristic intron-exon boundaries which would be expected if the deleted region constituted an optional exon that was subject to differential splicing. However, we cannot discount the possibility that the sequence of the legumin gene from which pDUB3 was derived encodes the necessary splicing sites which could effect the deletion of the sequences missing from that cDNA.

On balance, the data suggest that the absence of the repeats from the pDUB3 cDNA is an artefact. The fact that the deleted region is bounded by identical hexanucleotide sequences, GGCAGC (see Fig.10) suggests that the deletion may have arisen from a recombination event in the *E.coli* host used for cloning. Alternatively, transient base-pairing during cDNA synthesis may have effected the loss of an internal portion of the mRNA sequence.

Restriction endonuclease analyses of the legumin clones isolated from the cDNA library indicated that many (13 out of a total of 23) were similar to pDUB3 in that they also appeared to lack the repeats (see section 3.2.4.). This finding appears, initially, to be inconsistent with the hypothesis that the absence of the repeats from pDUB3 is an artefact. However, the finding that all the new pDUB3-like clones appeared to contain inserts of the same length as the pDUB3 insert suggests that they may not be independent clones, but may, instead, have arisen from the inadvertent contamination by pDUB3 of the cDNA or vector (pBR322) DNA samples used for the construction of the present library. Sequencing of the 5' and 3' termini of one or more of the new pDUB3-like cDNA inserts should indicate whether they are indeed identical to pDUB3 as is suggested by the restriction mapping data, or whether they are independent clones in which case sequence differences should be evident (at least) in the lengths of their poly(A) tails.

By analogy with the legumin genes, Hu *et al.* (1982) have previously reported the occurrence in a zein genomic clone of a 96bp tandem duplication which is absent from two otherwise extensively homologous cDNAs. These workers noted the presence of consensus splicing sequences near the junctions of each duplicated sequence, but also pointed out that splicing at either of these putative sites would change the translational reading frame of the gene leading to premature termination of protein synthesis. Thus, it remains unclear whether the observed differences between the genomic and cDNA zein clones are genuine or artefactual.

An unexpected result from the genomic blot was that a number of genomic fragments which hybridised weakly to the acidic and basic subunit probes hybridised relatively strongly to the probe for the repeats, and in addition, certain fragments hybridised exclusively, albeit weakly, to this probe (see Fig.13). This may be due partly to the fact that the G + C content of the repeats probe (55%) was higher than that of the probes for both the acidics (45%) and the basics (41%). Thus although the hybridisation filters were washed under identical conditions, the filter screened with the repeats probe was effectively washed at a lower stringency. A more intriguing possibility is that sequences homologous to the repeats are present in other possibly nonlegumin, genes. The relatively strong hybridisation of the 4.4Kb HindIII fragment to the repeats probe (see Fig.13) also raises the interesting possibility that a legumin gene containing more than three repeats is borne on that fragment.

The 6.5 Kb EcoRI genomic fragment which hybridised exclusively to the probe for the repeats probably corresponds to the 6.4 Kb EcoRI fragment present in the λLEG3 genomic clone produced by Shirsat (1984). A genomic fragment of similar size was shown to hybridise to a labelled, 1.8 Kb λLEG3 subfragment carrying the legumin gene sequence (Shirsat, 1984). The 6.4 Kb genomic fragment was not apparently detected by Croy *et al.* (1982), and was similarly not detected in the present work by probes for the acidic and basic subunits.

A noteworthy feature of the genomic hybridisations is that the cumulative hybridisation intensity of the bands in the HindIII or HindIII/BamHI tracks is apparently greater than that in the EcoRI tracks. This may reflect the inefficiency with which the relatively large EcoRI legumin fragments were transferred to the nitrocellulose filter compared to the transfer of the smaller HindIII fragments. Thus the three EcoRI fragments, each of >7 Kb, which appear to hybridize with an intensity of ∿1 gene equivalent (Croy *et al.*, 1982), may in reality carry more than one legumin gene. Indeed, as previously noted, the analysis of genomic clones has indicated the presence of a legumin pseudogene in addition to a normal legumin gene on the ∿13 Kb fragment (R.Croy, pers.commun.). Tne analysis of the smaller fragments in the HindIII and HindIII/BamHI tracks should therefore provide a more accurate estimation of the legumin gene copy number. Comparing the intensities of the bands in these tracks to the intensity of the 4.2 Kb EcoRI fragment suggests the presence of at least seven legumin genes in the pea genome which is consistent with the estimate of Shirsat (1984).

4.3. Analysis of Vicilin cDNAs.

Screening of the cDNA bank for vicilin cDNAs was complicated by the fact that very high backgrounds of nonspecific hybridisation were obtained in colony hybridisation experiments using the then available vicilin cDNA inserts from pDUB2 and pDUB4 (Lycett *et al.*, 1983a) as probes (see Fig. 4C). Similar anomalous results with these two probes have been independently obtained in this laboratory (R.Croy, J.Gatehouse pers.communs.), and both probes have been shown by Southern blot analysis to hybridise to pBR322 (Fig. 5), but the reasons for this are not understood. Contamination by vector sequences of the cDNA inserts recovered from agarose gels appears to be the most likely explanation since it has been ascertained by computer-assisted comparisons of DNA sequences that the pDUB2 and pDUB4 inserts do not share any significant homology with any region in pBR322 (G.Lycett, pers.commun.). The apparent inability of agarose gel electrophoresis to effect an absolute separation of DNA fragments has been noted previously (section 4.1). However, it is not clear whether the problem is inherent in the electrophoresis process or whether it is due to the presence in the restricted DNA sample of partially degraded vector fragments, some of which invariably match the size of, and hence comigrate with, the desired insert. Usually, the level of contamination is quite low (e.g. in the case of the legumin cDNAs) and does not interfere with colony hybridisation results (see Fig. 4B), but in the case of the pDUB2 and pDUB4 inserts, the contamination was, for unknown reasons, relatively very high (see Fig.4C). The problem was not encountered with all vicilin cDNAs since the inserts from the subsequently isolated vicilin clones pAD2-1 and pAD3-4, gave good results when used as probes in colony hybridisation assays (e.g. Fig.4D).

Vicilin cDNAs in the clone library were initially identified by Southern blot hybridisation to the pDUB2 and pDUB4 labelled inserts (see Fig.5). One of the cDNAs obtained was used to rescreen the library by colony hybridisation —— a total of 16 clones hybridised to that probe (Fig.4D; Table 7). A number of these clones were further characterised by hybrid-release translation (Fig.9), restriction mapping (Fig.7) and DNA sequencing (Fig.11).

Considerable homology, ∿85%, was found between coding regions of the sequenced vicilin clones. The cDNAs which encoded 50000-*M*r subunits were more closely homologous to each other than to a cDNA which encoded a 47000-*M*r subunit. This, and the fact that the cDNAs encoding 50000- and 47000-*M*r subunits could be distinguished by hybrid-selected translation experiments (see Fig.9), suggests the existence of subfamilies within the vicilin multigene family. It is noteworthy that the different vicilin cDNAs show more sequence variation than the legumin cDNAs which, excluding the deletion of the repeats in certain clones, are typically more than 98% homologous.

The cDNA which extended the furthest towards the 5' end of the vicilin mRNA, designated pAD3-4, encoded a 19 residue long signal peptide upstream from the N-terminus of a mature vicilin subunit (discussed in section 4.4.2.) plus 342 amino acids of a 47000-*M*r vicilin subunit,but lacked ∿266 bp of 3' coding sequence and all of the 3' untranslated region (see Fig.11). The 3' terminal 52 bp of pAD3-4 was completely homologous with the 5' end of another 47000-*M*r vicilin cDNA, pDUB4 (Lycett *et al.*,1983a); thus, these two clones are thought to be derived from transcripts of the same gene.

Another vicilin clone, pAD2-1, encoded two residues of the signal peptide, the entire sequence of a 50000-*M*r subunit, and 130 bp of the 3' untranslated region (Fig.11). There was almost complete homology between a large part of the pAD2-1 sequence and that of another sequenced cDNA, pAD7-13 : only 2 nucleotide substitutions were found in an overlapping region of 1360 bp (see Fig. 12). This indicates that the cDNA inserts in pAD2-1 and pAD7-13 were derived either from very similar (allelic?) genes or from a single gene, the observed mismatches being the results of inaccurate copying by reverse transcriptase which is known to be error-prone (Gopinathan *et al.*,1979). Outside the extensive regions of homologous sequences, significant sequence divergences were apparent at both the 5' and 3' termini of the pAD2-1 and 7-13 clones. The occurrence of these sequence differences does not contradict the suggestion that the two cDNAs were derived from the same gene since, as discussed below, the differences at the 5' ends of the clones probably reflect a cloning artefact in the pAD7-13 cDNA, whereas the sequence divergence at the 3' termini appears to have arisen from differential polyadenylation of the primary mRNA transcripts (discussed in section 4.3.1.).

The 430 bp sequence at the 5' end of the pAD7-13 cDNA consists of an inverted repeat (absent from pAD2-1) of an internal stretch of sequence (see Fig.12). Similar inverted sequences have been previously reported in numerous cloned cDNAs (e.g. Fagan *et al.*, 1980; Volckaert *et al.*, 1981; Weaver *et al.*,1981; Geraghty *et al.*, 1982; Rasmussen *et al.*, 1983). These inverse repeats are thought to be artefacts of cDNA cloning resulting either from the first cDNA strand "looping back" on itself during the reverse transcriptase reaction, or from the "slippage" of the hairpin loop during second strand synthesis. (The reader is referred to Fagan *et al.* (1980) for details of these mechanisms). The resemblance of the pAD7-13 inverse repeat to the previously reported examples, plus the fact that it contains seven in-phase stop codons and has no homology with the corresponding regions in other vicilin clones, strongly suggests that it is an artefact.

Restriction mapping and preliminary sequence data from another vicilin cDNA, pAD6-11, indicated that it too shared an extensive region of sequences identical to the pAD2-1 and pAD7-13 sequences, but that it probably also contained an artefactual 5' inverse repeat similar to,but different from, that seen in pAD7-13 (see Fig.7). No other vicilin cDNAs were sequenced, but restriction mapping analyses showed that a number of clones which hybridised to the ^{32}P-labelled pAD2-1 insert belonged to four additional classes. Thus, including the previously described pDUB2 and pDUB4 inserts (Lycett *et al.*,1983a), the presently characterised vicilin cDNAs (represented by pDUB2, pAD3-4, 2-1, 5-4, 5-5, 7-3 and 7-5) appear to be derived from at least seven distinct vicilin genes. When these cDNAs are aligned on the basis of common restriction sites (see Fig.7), one of them pAD7-3, appears to extend further than any of the sequenced clones towards the 5' end of the vicilin mRNA. If this preliminary observation is confirmed by more detailed restriction mapping, then the sequencing of that clone should give more information on the structure of the vicilin signal peptide (see section 4.4.2.), and the 5' end of the vicilin mRNA.

4.3.1. Structural Polymorphism in the 3' Untranslated Regions of Vicilin cDNAs.

The 3' noncoding sequence of pAD2-1 is identical to that of pAD7-13 except that it extends into a 31 bp A+T-rich region beyond the point at which a poly(A) tail is attached in pAD7-13 (see Fig. 11). The pAD2-1 cDNA apparently lacks a poly(A) sequence, but since the mRNA template was isolated on an oligo(dT) column, and second strand cDNA synthesis was primed by oligo(dT), a poly(A) tail must originally have been present downstream from the cloned 3' terminus. The differences between the 3' noncoding sequences of the two cDNAs, therefore, strongly suggest that alternative poly(A) addition sites were used to terminate the transcripts of the same, or two very similar, vicilin genes. Evidence for differential polyadenylation of primary transcripts has been previously found in zein cDNAs (Heidecker and Messing, 1983), hordein cDNAs (Rasmussen *et al.*,1983), *Agrobacterium* T-DNA genes (Dhaese *et al.*, 1983), and in a number of animal cDNAs (e.g. Tosi *et al.*, 1981; Setzer *et al.*, 1982, Early *et al.*,1980). The function of this mRNA heterogeneity is not clear in most cases (as in this work), where the use of alternative poly(A) addition sites affects only the length of the 3' untranslated region and not the encoded protein, unlike the situation where one of the potential poly(A) sites occurs within the gene coding sequence and differential polyadenylation leads to the production of functionally different proteins, e.g. the membrane and secreted forms of immunoglobulin μ chains (Early *et al.*,1980). Structural polymorphism in the 3' untranslated region might conceivably affect transport of the processed mRNAs from the nucleus to the cytoplasm (Setzer *et al.*,1982) and may have a role in the regulation of gene expression by affecting the stability of the transcripts.

Since the polyadenylation of certain vicilin mRNAs may occur at alternative poly(A) sites, it might be instructive to examine the distribution of putative polyadenylation signal sequences in the 3' noncoding regions of the vicilin clones. Various studies (Fitzgerald and Shenk, 1981; Montell *et al.*,1983; Higgs *et al.*, 1983) have shown that the highly conserved sequence AAUAAA, found 11-30 nucleotides upstream from the poly(A) tail in most animal mRNAs (Proudfoot,1982) is essential for the formation of the mature 3' end of the message prior to polyadenylation. However, sequence data from plant genes suggest that the polyadenylation signal sequences in plant mRNAs are more variable than their animal counterparts,, both with respect to the actual sequences involved and in their distance from the polyadenylation site. Reported variations include point-mutated homologues of the AAUAAA sequence, and overlapping or separate repeats of the archetypal sequence, and many plant genes appear to have the normal AATAAA sequence close to the stop codon while a variant of it occurs in the more usual position ∿20 bp upstream from the poly(A) tail (Lycett *et al.*,1983b; Messing *et al.*,1983; Dhaese *et al.*,1983). Not surprisingly, several putative polyadenylation signals can be identified in the 3' noncoding regions of the available vicilin cDNAs.

An AATAAA sequence precedes the poly(A) tail by 21 bp in the pAD7-13 cDNA (Fig.11) and is probably responsible for the selection of the polyadenylation site in that clone. However, it may not be the major signal sequence since the pAD2-1 cDNA is apparently polyadenylated at a site further downstream. In both pAD7-13 and pAD2-1, the AATAAA sequence overlaps with a variant of the consensus sequence to give AATGAATAAA which is similar to the minor polyadenylation signal sequence, AATGAATATA, in the octopine synthase gene (Dhaese *et al.*,1983). The pDUB2 cDNA also has an AATAAA sequence in the same relative position, but it is not known whether that sequence functions as a polyadenylation signal since the cloned cDNA terminates only 13 bp further downstream and does not have a poly(A) tail. The composite sequence, AATGAATAAA, present in pAD2-1 and pAD7-13 is not conserved in pDUB2 where the first T in that sequence is substituted by a G. The extent to which the selection of alternative polyadenylation sites is actually influenced by these putative signals is not known, but in the case of pAD2-1 and pAD7-13, it seems likely that the primary transcripts from which these clones were derived contained an additional signal sequence downstream from the point at which the pAD7-13 mRNA was polyadenylated. It may be worthwhile to sequence the 3' terminus of the pAD6-11 insert, in the event that it was polyadenylated at the same site as in pAD2-1, to see whether any additional 3'-proximal signal sequences are indeed present.

The 3' untranslated sequences of the vicilin cDNAs are less conserved than the coding sequences. This contrasts with the finding that the 3' noncoding regions of the conglycinin (soybean vicilin) genes are more conserved than the coding regions (Schuler *et al.*,1982b), but is similar to the relative degrees of sequence conservation found in the genes of other closely related proteins including the globin genes of different species (Efstratiadis *et al.*, 1980), the actin genes of *Drosophila* (Fyrberg *et al.*, 1981), and the insulin genes of various species (Sorokin *et al.*, 1982). To explain their unusual observation, Schuler *et al.* (1982b) suggested that the requirement for a particular secondary structure in the 3' noncoding regions of the conglycinin mRNAs may have constrained their divergence. However, similar constraints do not appear to have been imposed on the 3' untranslated region in pea vicilin mRNAs.

4.3.2. Amino Acid Sequences Predicted from Vicilin cDNAs.

Fig.26 shows a comparison of protein sequences predicted from the vicilin cDNAs with partial and complete amino acid sequences determined from purified vicilin subunits (protein sequence data from Lycett *et al.*,1983a). The extensive homology evident among the vicilin cDNAs is reflected by a high degree (∿80%) of amino acid

```
pAD3-4    IKPLMLLAI  AFLASVCVSS RSDQENPFIF KSNRFQTLYE NENGHIRLLQ KFDKRSKIFE NLQNYRLLEY KSKPHTLFLP QYTDADFILV
pAD2-1             .. ...PQ.S... ...K....F. .......... ...Q...... .......... ......I... .H....Y...
α                     .....      ........FQ .......... ...Q...... .......... .....   .. .QN...
α+β                   ▌......... .......... .......... .......... .......... ....R   .. N.
                                         F          V          Q  D
α+β+γ                  .........  .   ...... .......           ... ......

pAD3-4    VLSGKATLTV LKSNDRNSFN LERGDAIKLP AGSIAYFANR DDNEEPRVLD LAIPVNKPGQ LQSFLLSGTQ NQKSSLSGFS KNILEAAFNT
pAD2-1    ......I... ..PD...... .....T.... ..T...LV.. .....L.... ......R... ........N. ..QNY..... ......S...
α           ........ ..    .... .......... ..T.. L... ....DL.... ......R... ..                    .........
                                                                                                   S
α+β         ........ .......... .......... ..T...L... ....DL.... .......... ..Q....... ..    .... ..........
                             Y                                         R  N    L                   S
α+β+γ          .F... .LP...... ...     .. ...... ....   ... ......R...                             .....S...

                                α ↓ β
pAD3-4    NYEEIEKVLL EQQEQEPQHR RSLKDRRQEI NEENVIVKVS RDQIEELSKN AKSSSKKSVS SESGPFNLRS RNPIYSNKFG KFFEITPEKN
pAD2-1    D......... .EH.K.T... .....K..QS Q.......L. .G........ ...T...G.. ...E...... .G.....E.. ..........
pDUB2     DNA....I.. .EH.K.TH.. .G.R.K..QS Q.K....... KK........ .......... .R.E....K. SD.....QY. .......K..
α/β       ...  ....  .             ▌.....L SN........ .R........ .....R.... .......... .D.....NS. ..........
                                                       E                    N          S
α+β       .......... ..H....... .G.R....QS Q.K..   .. .E........ .....RR... .......... .......NY. ..........
                           TH G                                                        SD
α+β+γ     D       ... ..                                                ... ...E....K. SD....
                                                                            K

                                                                        β ↓ γ
pAD3-4    QQLQDLDIFV NSVDIKVGSL LLPNYNSRAI VIVTVNEGKG DFELVGQRNE NQ---GKEN DKEEEQ-EEET SKQVQLYRAK LSPGDVFVIP
pAD2-1    P......... ...E..E... ...H...... .......... .......... ..QEQR..D .E....G...I N....N.K.. ..S.......
pDUB2     P......... .Y.E..E... W..H...... .......... .......... ..QGLRE.D .E....R.... KN...S.K.. .T........
β/γ       ..I..I.... ......E... .I........ LVIW...... .......... ..---...▌ H....-..... .......... ..........
α+β/γ     .......... ..D   E... ........          ... .......... ..---...▌      -..... .......... ..........
α+β+γ                      E... ...H....

pAD3-4    AGHPVAINAS SDLNLIGFGI NAENNERNFL AGEEDNVISQ VERPVKELAF PGSSHEVDRL
pAD2-1    ......LK.. .N.D.L.... .....Q.... ..D....... IQ........ ...AQ....I LENQKQSHFA DAQPQQRERG SRETRDRLSS
pDUB2     ......VR.. .N...L.... .......... .......... IQKQ..D.T. ...AQ..... .......Y.. N.......TR .Q.IKEHLY.
γ         ......?... .......... .......... .......... .......... .......... .......Y.. N...L...TR .Q.▌
                                                                                 L A
α+β+γ            R.. .N.D...... ..     ... ..D            D... ....Q....                      .TR

pAD2-1    V
pDUB2     ILGAF
```

Figure 26. Comparisons of protein sequences preducted from the vicilin cDNAs (see Fig.11) with partial and complete amino acid sequences determined directly from purified vicilin subunits (α,β,γ,α+β, α+β+γ — see Fig. 27 for the derivation of these subunits. The direct amino acid sequence data are taken from Lycett *et al.*(1983a). The sequence labelled pAD3-4 comprises a composite pAD3-4/pDUB4 - derived amino acid sequence.

▌ indicates that the N- or C-terminus was determined directly. Other symbols are as used in Fig. 25. In place of the boxed H residue down stream from the β:γ cleavage site in the γ-subunit, Spencer *et al.* (1984) determined a K residue which matches the sequence predicted by the pAD3-4 cDNA. A potential glycosylation site encoded by pAD3-4 is overlined.

conservation in the overlapping regions of the predicted protein sequences. The amino acid substitutions which occur are generally dispersed throughout the sequences, but there are three "hot spots" within which sequence divergencies are concentrated.

Two occur in the vicinities of potential proteolytic cleavage sites and are discussed in section 4.4.2. The other occurs in the C-terminal regions encoded by pAD2-1 and pDUB2 (N.B. the pAD3-4/ pDUB4 sequence does not extend into this region). The protein sequence encoded by pDUB2 is terminated at a phe residue by a pair of tandem stop codons (see Fig.11). Though the identical stop codons are present in pAD2-1, the encoded protein ends at a val residue four residues upstream from the predicted, pDUB2-encoded C-terminus of the protein. The 17 amino acids immediately preceding the tandem stop codons in pAD2-1 and pDUB2 contain 13 (77%) mismatches (see Fig.26). Direct amino acid sequencing has identified the glu residue which occurs thirteen residues upstream from the paired stop codons as the C-terminus of the 16000-*M*r (γ) vicilin subunit (Lycett *et al.*,1983a; Fig.26). This suggests that C-terminal extensions may be removed post-translationally from some vicilin precursors, assuming that the mature products of the pDUB2- and pAD2-1- encoded subunits have the same C-termini as the 16000-*M*r subunit.

The protein sequence deduced from the pAD3-4 cDNA (Fig.26) contains an N-A-S sequence which is a potential site for N-glycosylation (Struck *et al.*,1978). The failure of direct protein sequencing to identify the residue occurring in the 16000-*M*r subunit in the position of the predicted asn (see Fig.26) suggests that this site is glycosylated *in vivo*. Potential glycosylation sites are completely absent from the other predicted protein sequences.

4.4. Proteolytic Processing of Pea Storage Proteins.

4.4.1. Legumin.

None of the available cDNAs extend far enough towards the N-terminus of the legumin subunit precursor to show the existence of a signal peptide. However, the protein sequence deduced from a legumin genomic clone (Lycett *et al.*,1984a) showed that a 21 residue-long peptide which has all the characteristics of a signal peptide (see von Heijne, 1983) precedes the N-terminus of the mature protein. The presence of signal peptides has been demonstrated in almost all the studied seed storage proteins, and there is evidence that they are removed co-translationally as the polypeptides are sequestered into the endoplasmic reticulum (reviewed by Gatehouse *et al.*,1984; Higgins, 1984).

Legumin polypeptide precursors also undergo post-translational endo-proteolytic cleavage to generate the acidic (α-) and basic (β-) subunits (Croy *et al.*,1980a; Spencer and Higgins, 1980). The amino acid sequences predicted by pAD4-4 and pAD10-5 in the vicinity of the cleavage site are identical to the previously published pDUB3/1 - encoded sequences (Croy *et al.*,1982; see Fig.25). Croy *et al.*(1982) speculated that cleavage might occur at a pair of adjacent arginine residues five amino acids upstream from the N-terminus of the basic subunit by analogy with the processing of animal hormone precursors. If that hypothesis is correct, then further peptidolytic processing would be required to generate the N-terminus of the mature β-subunit (Croy *et al.*,1984). A similar mechanism involving the removal of a short linker between the acidic and basic subunits of glycinin (soybean legumin) has subsequently been proposed (Nielsen, 1984). However, the recent determination of the C-terminal sequence of the pea legumin acidic subunit (Lycett *et al.*,1984b; see Fig.25) shows that it extends to the asparagine residue adjacent to the N-terminus of the basic subunit. This data indicates that post-translational cleavage of legumin precursors occurs at only a single site on the C-terminal side of the asparagine residue within the sequence R-R-Q-G-D-N-/G-L-E-E-T, and probably not at the paired R-R residues.

4.4.2. Vicilin.

Data from various *in vitro* translation studies have previously shown that vicilin precursor polypeptides contain a signal peptide at their N-terminus which, as with legumin, is removed during translocation of the protein across the endoplasmic reticulum membrane (reviewed by Gatehouse *et al.*,1984). Consistent with this data, the cDNA in pAD3-4 encodes a signal peptide-like sequence (von Heijne, 1983) of 19 amino acid residues at its 5' terminus

(Fig.26). A predicted methionine occurs 15 residues upstream from the N-terminus of the mature subunit, but it is not clear whether that methionine constitutes the initiation codon or whether translation of the pAD3-4 mRNA is initiated at another methionine codon further upstream from the 5' terminus of the cloned cDNA. By way of comparison, phaseolin (French bean vicilin) contains a signal peptide of 21-26 residues (Slightom *et al.*,1983).

The possibility that vicilin subunits may undergo C-terminal proteolytic processing was noted earlier (section 4.3.2.). Other plant proteins which are subject to C-terminal proteolysis include thaumatin (Edens *et al.*,1982) and pea lectin (Higgins *et al.*,1983a). A 6 residue-long C-terminal peptide is cleaved from the thaumatin precursor. This peptide is predominantly acidic, in contrast to the overall, highly basic character of thaumatin, and may be involved in the compartmentalization and ultimately in the stability of the protein (Edens *et al.*,1982; Edens *et al.*,1984). Four C-terminal residues also appear to be cleaved from the pea lectin. The functions of the C-terminal extensions in the vicilins and the lectin are unknown, but the occurrence of C-terminal processing of these proteins is consistent with the localization of a carboxypeptidase in the seed protein bodies of *Vigna radiata* (Harris and Chrispeels, 1975), assuming that the terminal residues are removed sequentially rather than as an oligopeptide.

In addition to the removal of the N-terminal signal peptide and the possible removal of C-terminal extensions, vicilin subunits of *M*r ∿50000 are also subject to extensive endoproteolytic processing. Unlike legumin where cleavage of all precursors appears to be obligatory, only certain ∿50000- *M*r subunits undergo cleavage. From comparisons of amino acid sequences predicted by vicilin cDNAs with direct protein sequence data, Gatehouse *et al.* (1982) and Spencer *et al.* (1983) independently presented a model which accounts for the derivation of vicilin polypeptides of *M*r <50000 by post-translational proteolysis of susceptible ∿50000-*M*r subunits (see Fig.27). This model envisages the presence of up to two potential cleavage sites, designated the α:β and β:γ sites (Gatehouse *et al.*,1982) in the 50000-*M*r vicilin precursors and all the smaller vicilin subunits (α,β,γ,α+β and β+γ) are generated by proteolysis at one or both of these sites.

FIGURE 27. Derivation of vicilin subunits from precursors of ∿50000-*M*r. The SDS-polyacrylamide gel (on the right) shows the major vicilin subunits in the vicilin fraction isolated from mature pea seeds. (From Gatehouse et al., 1984).

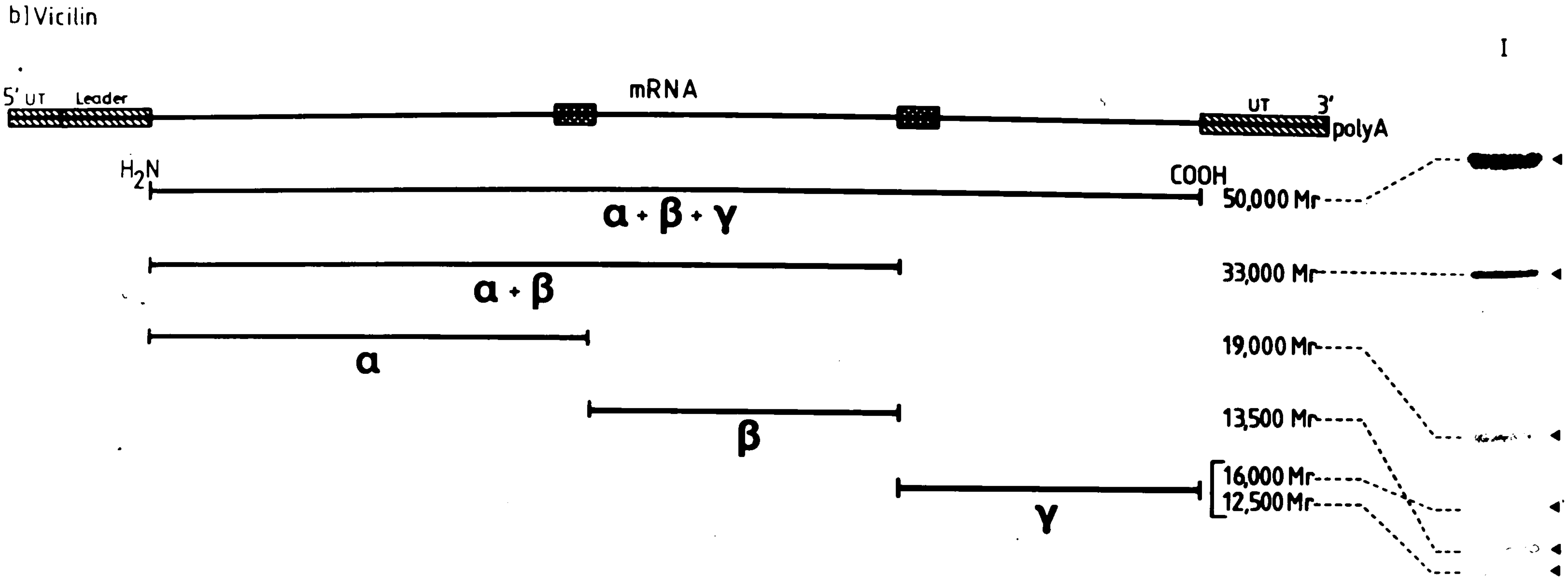

b) Vicilin
5' UT
Leader
mRNA
UT
3'
polyA
I
H_2N
COOH
α + β + γ
α + β
α
β
γ
50,000 Mr
33,000 Mr
19,000 Mr
13,500 Mr
16,000 Mr
12,500 Mr

Fig. 27

The availability of the new cDNA clones, pAD3-4 and pAD2-1, allows more extensive comparisons of sequences in the regions of potential proteolytic cleavage. The pAD3-4 cDNA is particularly important in this regard since it encodes a 47000-*M*r precursor that is known to be proteolytically cleaved *in vivo* at the β:γ site to give a 33000-*M*r (α+β) subunit as one of the products (Gatehouse *et al.*, 1981). Protein sequences derived from the available vicilin cDNAs have been compared with the determined partial sequences of vicilin subunits in Fig.26. The sequence predicted by pDUB2 is almost idential (only one lys → arg substitution occurs in a 20 residue-long sequence) to the sequence of the 33000-*M*r (α+β) subunit in the region spanning the α:β cleavage site, but shows significantly less homology with the N-terminal sequence of a 19000-*M*r (β) subunit. This suggests that the pDUB2-encoded product is not susceptible to cleavage at the α:β site. Similarly, the pAD3-4 and pAD2-1 encoded polypeptides would also not be expected to undergo cleavage at the α:β site since although the protein sequences predicted by these cDNAs show less homology to the 33000-*M*r subunit sequence, the divergencies are conservative ones : e.g. glycine and arginine residues are substituted by functionally similar serine and lysine residues respectively.

The region in the vicinity of the potential β:γ cleavage site shows marked differences in the amino acid sequences predicted by the various cDNAs. In this region, the amino acid sequence derived from pAD3-4 is completely homologous to the determined terminal sequences of the subunits produced by cleavage at that site (i.e. the β or α+β subunits, and the γ- subunit), but differs from the sequence of an uncleaved 50000-*M*r subunit. This perfect matching of the predicted and determined amino acid sequences is consistent with the proposed cleavage of the pAD3-4-encoded subunit at the β:γ site (Gatehouse *et al.*, 1981).

The predicted sequence of the pDUB2- encoded subunit contains an extra four amino acids relative to the pAD3-4-encoded product in the immediate vicinity of the β:γ site, and matches the partially determined sequence of an uncleaved 50000-*M*r subunit which contains at least three of these four extra residues (see Fig.26). On that basis, the pDUB2 subunit would not be expected to undergo cleavage at the β:γ site. The protein sequence deduced from pAD2-1 resembles the pDUB2-derived sequence to the extent that it also contains four extra amino acid residues relative to the cleaved precursor, but these extra residues are not homologous to the partially determined sequence of the uncleaved subunit. The twenty amino acids downstream from the potential β:γ site in the pAD2-1-encoded protein contain six mismatches (including a one-residue insertion) when compared with the N-terminal sequence of the γ-subunit. Unfortunately , no protein sequence data are available for this region in the uncleaved precursor, but the significant degree of sequence divergence with the γ-subunit tends to suggest that the pAD2-1-encoded subunit is also not susceptible to cleavage at the β:γ site.

4.4.3. Sequence Specificity of the Endo-proteolytic Cleavage Sites in Seed Proteins.

The protein sequences derived from the legumin and vicilin cDNAs give only limited clues to the sequence specificity of the endo-proteolytic processing sites since only two precursors which undergo cleavage (i.e. legumin and 47000-*M*r vicilin) have been identified. However, certain members of another group of seed proteins, the lectins, are also susceptible to endo-proteolytic processing in a manner apparently analagous to the major storage proteins (Chrispeels, 1984; Higgins *et al.*,1983b). The lectins from pea, lentil and *Vicia faba* are synthesised as high molecular weight precursors which are cleaved post-translationally in the protein bodies to produce two subunits of ∿17000- and ∿6000-*M*r, whereas other lectins including those from *Phaseolus vulgaris*, soybean, sainfoin, jackbean and *Dioclea grandiflora* are not subject to the same form of processing.

To try and determine the sequence specificity of the endo-proteolytic cleavage, the sequences in the vicinity of the potential cleavage sites in legumin and the vicilins, as well as glycinin and several lectins are compared in Table 9. Examination of the sequences allows two generalisations to be made : (i) the cleavage sites occur within highly hydrophilic regions of the proteins; and (ii) cleavage occurs on the C-terminal side of an asparagine residue. A consensus sequence, $\mathrm{N}\text{-}/\mathrm{X}_1\text{-}\mathrm{X}_2\text{-}{\mathrm{D} \atop \mathrm{E}}\text{-}\ \mathrm{E}$ in the vicinity of the cleavage site may be formulated —— in three

TABLE 9 Potential Proteolytic Cleavage Sites in Legume Seed Proteins

Protein	Sequence of precursor in vicinity of potential cleavage site.	Cleavage?	Reference
Storage Proteins			
Legumin	K G K S R R Q G D N ↓ G L E E T V C T	Yes	Lycett *et al.*(1984b)
Glycinin	Q R Q S K R X X - N ↓ G I D E T I C T	Yes	Nielsen (1984)
Vicilin (pDUB2)α:β	K E T H H R R G L R ↓ D K R Q Q S Q E	No	Lycett *et al.*(1983a)
Vicilin (pAD2-1)α:β	K E T Q H R R S L K ↓ D K R Q Q S Q E	No	This work
Vicilin (pAD3-4)α:β	Q E P Q H R R S L K ↓ D R R Q E I N E	No	This work
Vicilin (pDUB2) β:γ	E N Q Q G L R E E D ↓ D E E E E Q R E	No	Lycett *et al*(1983a)
Vicilin (pAD2-1)β:γ	E N Q Q E Q R K E D ↓ D E E E E Q G E	No	This work
Vicilin (pAD3-4)β:γ	E N Q ——— G K E N ↓ D K E E E Q — E	Yes	This work
Lectins.			
pea	V L T V S L T Y P N ↓ S L E E E N ↓ V T S Y T L	Yes	Higgins *et al.*(1983a)
V.faba[(a)]	V L S V T L L Y P N L T G Y T L	Yes	[Hemperly *et al.*(1979) Hopp *et al.*, (1982)]
lentil	L Q N G V T S Y T L	Yes	Foriers *et al.*, (1981)
P.vulgaris	V F S V S L S N P — S T ——— G K S N N V	No	Hoffman *et al.*(1982)
Soybean	L L V A S L V Y P — S Q ——— R T S N I L	No	Vodkin *et al.*(1983)
sainfoin[(a)]	V S S F Y R N K P D ——— D I F T V	No	Kouchalakos *et al*(1984)
jackbean[(a)]	R L S A V V S Y P N ——— A D A T S V	No	[Wang *et al.*(1975) Cunningham *et al.*(1975)]
D.grandiflora[(a)]	R L S A V V S Y S G ↓ ——— S S S T T V	No	Richardson *et al.*(1984)

a. Only amino acid sequence data for the mature protein (as opposed to mRNA sequence data for the precursors) are available. = carboxypeptidase activity; ↓ = cleaved proleolytic sites; = uncleaved proteolytic site. Gaps(——) have been inserted in the sequences to mamimize homology.

of the four cases for which full sequence data are available for the cleaved precursors, X_1 is a small, neutral residue (G or S) while X_2 is a hydrophobic residue (L or I). However, as discussed later, it is likely that this sequence may be involved in making the asparaginyl bond generally accessible to the peptidase, rather than being specifically required for enzyme activity.

In pea lectin (Higgins *et al.*,1983a) and in glycinin (Nielsen, 1984), endoproteolytic cleavage at the specified asparagine residue appears to be accompanied by further peptidolytic processing to generate the termini of the mature subunits. Similar processing is thought to occur after the primary cleavage reaction in the lentil and *V.faba* lectins (Foriers *et al.*,1981), and since only protein sequence data for the mature subunits are available, as opposed to mRNA sequence data for the precursors, it is not possible to confirm whether the initial cleavage occurs within the sequence specified above.

The precursors of castor bean lectins (agglutinin and ricin) are also cleaved to produce two subunits. However the cleavage sites in the prolectins (Lamb, 1984) do not appear to be located within an obviously hydrophilic region as in the other proteins discussed. Moreover, the agglutinin and ricin sequences (Lamb,1984) have little homology with the other lectin sequences (and have thus been omitted from Table 9), and it is possible that the mechanism of their processing varies somewhat from that observed in other species. Nevertheless, an analysis of the sequences of proagglutinin and proricin suggests that cleavage may occur initially at an asparagine residue (within the sequence N-/A-D-V-C),followed by carboxypeptidolytic processing to remove a 12 residue-long linker between the mature subunits.

The lectins from jackbean and *Dioclea grandiflora* undergo endoproteolytic processing at different sites from those in the pea, lentil and *V.faba* lectins. It is noteworthy that in these cases too, cleavage apparently occurs at asparagine residues within the sequences N-/S-T-H-Q-$\underline{T}$ (jackbean; Wang *et al.*,1975) and N-/S-I-A-D-$\underset{E}{\underline{A}}$ (*D.grandiflora;* Richardson *et al.*,1984).

In general, the sequences of the proteins which are not susceptible to endo-proteolytic cleavage (see Table 9) are completely divergent in the immediate vicinity of the potential cleavage site.

It has previously been suggested (Gatehouse *et al.*,1983) that the presence of a functionally distinct, neutral-hydrophobic-basic residue sequence preceding the potential processing site in vicilin precursors was not conducive to cleavage. That hypothesis was based on the occurrence of G-L-R or S-L-K sequences preceding the uncleaved α:β site in the pAD3-4- and pDUB2-derived protein sequences and a G-L-R sequence preceding the uncleaved β:γ site in the pDUB2-encoded subunit (see Table 9). However, it does not seem to be generally applicable since no similar cleavage-inhibitory sequences can be identified in the region preceding the uncleaved β:γ site in the pAD2-1-encoded vicilin subunit, nor in any of the uncleaved lectins.

Even if the postulated N-/X-X-$\genfrac{}{}{0pt}{}{D}{E}$-E consensus sequence proves to be an accurate formulation of the sequence recognized by the processing enzyme system, the mere presence of that sequence in a seed protein will probably not ensure cleavage. Presumably, other factors, such as the accessibility of the site and the three-dimensional protein structure in the immediate vicinity are also likely to be important. By analogy, although the tripeptide sequence N-X-$\genfrac{}{}{0pt}{}{T}{S}$ specifies the glycosylation of asparagine residues (Struck *et al.*,1978), relatively few of these sequences in proteins are actually glycosylated (Marshall, 1972). Instead, the residues which do undergo glycosylation have been found to be associated with β-turn conformations and/or hydrophilic regions of the protein (Aubert and Loucheux-Lefebvre, 1976; Aubert *et al.*,1976). The glycosylation of these specific asparagine residues is thought to reflect the fact that both β-turns and hydrophilic regions are typically associated with the surfaces of globular proteins (Beeley, 1977; Struck *et al.*,1978). Interestingly, the secondary structures predicted for the legumin and vicilin precursors (Croy and Gatehouse, 1985) indicate that the cleavage sites in these proteins are also associated with β-turn conformations occurring between regions of more rigidly defined structure (α-helix or β-sheet), and it may be that this geometrically well defined domain facilitates the interaction with the processing enzyme. It may be tentatively concluded that the endoproleolytic,processing enzyme localized in the protein bodies specifically cleaves at asparagine residues, and the susceptibility of particular asparaginyl bonds results from their being situated in accessible regions (typically β-turn conformations in hydrophilic regions) of the protein.

4.4.4. Functions of the Proteolytic Processing of Pea Storage Proteins.

The functions of the endoproteolytic processing of legumin and vicilin precursors are unclear. Proteolysis of legumin (and all other 11S globulins studied —— Higgins, 1984) appears to be obligatory and may affect the molecular structure of the protein in a way that is vital for its proper packaging and deposition, or mobilisation during seed germination.

Post-translational cleavage of vicilin subunits may have similar functions. However, since uncleaved 50,000-*M*r vicilin subunits are the most abundant size class in the native vicilin fraction of pea, processing is not absolutely essential for the storage and metabolism of the vicilin proteins. Consistent with this view is the fact that the vicilins of *P.vulgaris* (Slightom *et al.,* 1983) and *G.max* (Beachy *et al.,* 1981) do not undergo significant post-translational proteolysis. It is possible that in those vicilins which are processed, cleavage did not originally serve any useful function, and instead merely reflected the existence of sites that were accessible to an asparagine-specific protease present in the protein bodies. It may be noted that the location of these sites within β-turn conformations on the surface of the protein minimizes the disruption of secondary and tertiary structure that may result from cleavage at these sites. However, this initially fortuitous processing may have affected the speed with which the proteins were mobilised at the onset of germination. Thus, the varying degrees to which the 7S proteins are proteolytically processed in different species may reflect differences in the timing and rate at which storage reserves need to be mobilised during germination (Lycett *et al.,* 1983a).

4.5. Homology of Pea Storage Proteins to Other Legume Storage Proteins.

Despite the frequent absence of immunological cross-reactions, the physico-chemical properties of the 11S and 7S proteins from various legumes have long suggested that these proteins are very similar (Derbyshire *et al.*,1976; Gatehouse *et al.*,1984). The recent acquisition of DNA sequence data for legumin and glycinin genes has enabled more precise comparisons of the homology between these proteins. Using dot matrix comparisons of legumin and glycinin amino acid sequences deduced from their nucleotide sequences, Croy *et al.* (1984) have shown that the two proteins are highly homologous over their entire length except in the region of the tandem repeats in pea legumin. As noted previously, glycinin contains only a single copy of a sequence related to the legumin repeat sequences. Interestingly, legumin-type proteins have also been discovered in non-legume species such as oats (Brinegar and Peterson, 1982) and rice (Yamagata *et al.*,1982; Wen-Ming *et al.*, 1983).

The availability of nucleotide sequences for the 7S proteins of pea, soybean and French bean has confirmed that these proteins, too, are highly homologous over much of their length (Lycett *et al.*, 1983a; Croy *et al.*,1984). However, the vicilin sequence differs markedly from those of phaseolin and conglycinin in the regions of the endoproteolytic processing sites (Lycett *et al.*,1983a). This is significant since the latter two proteins do not undergo processing comparable to that of vicilin. It will be interesting to compare the (as yet unavailable) sequences in these regions of the vicilin from *V.faba* which is processed similarly to pea vicilin (Scholz *et al.*,1983). A partial-length cDNA clone for convicilin has shown that it too shares substantial homology with vicilin, but less homology with phaseolin or conglycinin (Casey *et al.*,1984).

4.6. Cloned cDNAs Encoding Seed Proteins Other than Legumin or Vicilin.

Twenty of the fifty-nine clones which hybridised to cotyledon poly(A)$^+$ RNA did not hybridise to cDNAs for legumin or vicilin proteins. The intensities of hybridisation of these clones to the mRNA probe were in most cases relatively weak and the cDNAs were not extensively characterised. Of the few which gave strong signals with the mRNA probe, the cDNAs from two clones, pAD9-2 and pAD6-2, were subjected to restriction mapping analyses since their sizes ($\sim$1850 bp) suggested that they were about the right length to code for a convicilin 71000-*M*r subunit. The derived restriction maps of the two cDNAs appeared to be very similar to each other (see Fig.8), but bore no significant resemblance to the map of the convicilin cDNA isolated by Domoney and Casey (1983). This does not preclude the possibility that pAD6-2 and pAD9-2 are convicilin cDNAs since there is more than one convicilin polypeptide (Croy *et al.*,1980b) and thus several convicilin genes in the pea genome, and as the data from the vicilin cDNAs illustrate, restriction maps of even quite closely homologous cDNAs may vary considerably.

Domoney and Casey (1984) have identified a cDNA which hybrid-selects mRNAs encoding an 80000-*M*r variant of a legumin precursor dubbed "big legumin". It is therefore also possible that the pAD9-2 and pAD9-2 cDNAs encode a "big legumin" subunit. Unfortunately, no restriction map is available for the 80000-*M*r legumin cDNA to allow comparisons with the map of pAD9-2.

Apart from the storage proteins, other relatively abundant seed proteins whose mRNAs may comprise a significant proportion of the total mRNA population include the pea lectin and a number of albumin proteins of *M*r $\sim$8000, 25000 and 100000 (Gatehouse *et al.*,1984). The length of the pAD9-2 cDNA indicates that it is too long to code for the lectin precursor (*M*r ≃25000—Higgins *et al.*,1983a) or the smaller albumins, but it could encode part of the 100000-*M*r albumin. It should be noted that the distribution of restriction sites in pAD9-2 (Fig.8) indicates the absence of duplicated sequences comparable to the artefactual repeat in pAD7-13, and it may therefore be assumed that the clone contains a genuinely long cDNA insert. Further characterisation of pAD9-2 by hybrid-selected translation and, if necessary, DNA sequencing should therefore be undertaken. It may also be worthwhile to characterise more fully other isolated cDNAs which hybridised strongly to the poly(A)$^+$ RNA probe but not to legumin or vicilin cDNA probes (e.g. pAD2-11, 3-1, 5-13 and 10-2; see Table 7). The pAD3-1 cDNA was not sized but the sizes of the other cDNAs mentioned ($\sim$0.8-1.0 Kb) suggest that they may encode either the pea lectin or the albumin proteins.

4.7. Expression of Vicilin Subunits in *E.coli*.

Various factors limit the isolation of certain subunit precursors (see the Introduction, section 1.5.). However, to study the processing of vicilin sununits in detail at the molecular level, pure samples of these precursors are essential, and to this end, several expression plasmids designed to express vicilin molecules in *E.coli* cells were constructed (for recent reviews on the expression of eukaryotic genes in *E.coli*, see Maniatis *et al.*, 1982; Harris, 1983; Gatenby, 1983). In these plasmids, the operator-promoter region of phage $\lambda(\lambda O_L P_L)$ was fused to various vicilin cDNA sequences via translation initiation signals derived from either the replicase gene of phage MS2 or the λ*c*II gene. Thermo-inducible regulation of the P_L promoter was effected by maintaining the plasmids in defective lysogens containing a temperature-sensitive mutation (λ*c*I857) in the phage repressor gene. Under inducing conditions, the levels of vicilin accumulated by the bacteria varied dramatically depending on the host cell used and, in particular, on the plasmid construction.

4.7.1. Effect of Host Strain on Vicilin Synthesis.

In situ immunoassay of various λ*c*I857 strains harbouring the expression plasmids indicated that the *E.coli* strains K12ΔH1Δ*trp* and N99λ*c*I857 were best suited for the expression of vicilin (Fig. 19). The use of a protease-deficient strain, SG4044(p*c*I857), did not lead to any apparent increase in vicilin yields. By contrast, Remaut *et al.*, (1983a) obtained a 100% increase in the accumulation of β-interferon in strain SG4044(p*c*I857) compared to K12ΔH1Δ*trp*. The present results suggest that proteolytic degradation of vicilin in the bacterial cells was not a major influence on the yields obtained.

Vicilin expression in K12ΔH1Δ*trp* and N99λ*c*I857 was also monitored by SDS-PAGE analysis of total bacterial cell extracts. In contrast to the results of the *in situ* immunoassays (Fig.19) which indicated that K12ΔH1Δ*trp* and N99λ*c*I857 cells were equally well suited to vicilin expression, much higher levels of vicilin expression were detectable in the K12ΔH1Δ*trp* cells as judged by the electrophoretic analysis (see Fig.21). In the preparation of the cell extracts, it was noticed that a relatively small mass of cells was recovered by centrifugation of the induced N99λ*c*I857 culture compared to the size of the pellet recovered from the induced culture or from induced and uninduced K12ΔH1Δ*trp* cells. This phenomenon was further investigated by monitoring the growth characteristics of N99λ*c*I857 and K12ΔH1Δ*trp* cells under non-inducing (30^oC) and inducing (42^oC) conditions. The $O.D._{650}$ of the N99λ*c*I857 culture increased normally at 30^oC but decreased sharply upon induction (Fig.23), which suggested that cell lysis was occurring under inducing conditions. Thus the apparently low levels of vicilin accumulation shown by SDS-PAGE analysis was probably due to the fact that a large proportion of the protein was released into the culture medium upon cell lysis. Such an occurrence would not have adversely affected the results of the *in situ* screening.

4.7.2. Vicilin Yields Obtained from Different Plasmid Constructions.

The vicilin expression plasmids pAD2-1.expl(+) and pAD2-1.exp2(+) are derived from the expression vectors pPLc24 (Remaut *et al.*,1981) and pPLc245 (Remaut *et al.*,1983a) respectively. Both these vectors contain the $\lambda O_L P_L$ region and the translation initiation signals of the phage MS2 replicase gene (see section 2.1.3). In pAD2-1.expl(+), the vicilin cDNA is attached to the N-terminal 98 amino acids of MS2 replicase while in pAD2-1.exp2(+) the cDNA is attached directly to the initiation codon. A protein of *M*r ∿62000 which reacted with antivicilin IgG was accumulated to high levels (very approximately ∿5% of the total cell protein as judged by a purely visual inspection of the gel; see Fig.20) in induced K12ΔH1Δ*trp* cells transformed with pAD2-1.expl(+). The synthesis of this protein is consistent with the predicted *M*r of the MS2 replicase-vicilin fusion product. A negative control plasmid, pAD2-1.exp2(-), which contains the pAD2-1 insert cloned into pPLc24 in the opposite orientation to the direction of P_L- initiated transcription, did not synthesize any detectable vicilin (Figs.19 and 20). This confirms that in the exp(+) plasmids, the synthesis of vicilin is under the control of the P_L promoter.

By contrast, pAD2-1.exp2(+) failed to direct the synthesis of any detectable amounts of vicilin (Figs. 19,21 and 22). Faced with this discrepancy, it was necessary, first of all, to establish that pAD2-1.exp2(+) had been constructed correctly (i.e.maintaining the correct phasing of translational reading frames). Sequencing across the junction between the 5' terminus of the vicilin cDNA and the vector DNA (see Fig.15) revealed that the N-terminal codon of mature vicilin had been correctly ligated to the initiator codon. Furthermore, extensive restriction mapping did not reveal any gross molecular rearrangements in the plasmid. Since it was still possible that the operator-promoter region in pAD2-1.exp2(+) had suffered a small molecular rearrangement which was not detectable by restriction mapping, this region was replaced by the homologous region in pAD2-1.exp1(+) which was previously shown to be fully functional (see Fig.24). However the reconstructed plasmid still failed to produce any detectable vicilin. Thus, it may be concluded that the failure of pAD2-1.exp2(+) to direct the synthesis of vicilin was not due to faults in the construction of the plasmid. Alternative explanations will be considered in section 4.7.3.

The plasmids pAD2-1.exp3(+) and pAD2-1.exp4(+) were constructed from the expression vector pAS1 (Rosenberg *et al.*,1983) which contains the $\lambda O_L P_L$ region and translation initiation signals from the λcII gene (see section 2.1.3.). The pAD2-1.exp3(+)- encoded protein should contain at its N-terminus three residues encoded by the BamHI linker plus the two residues of the vicilin signal peptide present on the pAD2-1 cDNA insert. In pAD2-1.exp4(+) on the other hand, the linker and signal peptide residues were deleted by the combined activities of T4 DNA polymerase and mung-bean nuclease, so that the synthesised product should correspond to a mature vicilin subunit free of any extraneous sequences. Upon induction, both plasmids synthesised ∿50000-*M*r proteins which reacted with anti-vicilin (Fig.22). However, the levels of synthesis achieved was ∿3-5-fold lower than in cells harbouring pAD2-1.exp1(+) (see Table 8).

For the reasons previously discussed (section 4.4.2.), the proteins synthesised from the plasmids containing the pAD2-1 insert would not be expected to undergo cleavage in the pea seed. On the other hand, the cDNA insert in pAD3-4 encodes a vicilin subunit which is cleaved *in vivo* at the β:γ site, but some 265 bp of vicilin coding sequence is missing from its 3' terminus. To obtain an effectively full-length cDNA encoding a cleavable proteolytic site, the appropriate 3'-proximal fragments from the pDUB4 and pDUB2 cDNAs corresponding to the missing pAD3-4 region were added on to the 3' terminus of the pAD3-4 insert (Fig. 17). Though the resulting vicilin cDNA was assembled from three separate cDNA molecules, it may justifiably be regarded as a dihybrid molecule since the pAD3-4 and pDUB4 cDNAs are thought to be derived from the same gene. Thus, only ∿13% of the hybrid cDNA was contributed by a heterologous source (i.e. the pDUB2 cDNA).

The plasmid pAD3-4.exp2(+) comprised the pAD3-4 hybrid insert cloned into pAS1 (see section 2.1.3.). This plasmid directed the synthesis of a ∿47000-*M*r protein which reacted with antivicilin IgG (Figs. 21 and 22). However, in the construction of pAD3-4.exp1, it was evident that a strong selection pressure had been exerted against plasmids containing the hybrid insert in the appropriate orientation for expression. The ligation products formed between pAS1 and the hybrid insert were initially used to transform strain N99λ*c*I857. The resulting transformants all contained plasmids with the insert positioned in the opposite orientation to the direction of transcription, and it was eventually necessary to transform a *c*I$^+$ strain (producing a wild-type *c*I repressor) to obtain plasmids with the insert in the appropriate orientation for expression. The fact that plasmids containing the insert in both orientations could be isolated in roughly equal proportions from transformed *c*I$^+$ hosts indicates that the negative selection pressure did not operate at the level of the ligation of the insert to the vector. Rather, the evidence suggests that probably low-level expression of the vicilin cDNA in a *c*I857 host gave rise to a deleterious phenotype which killed, or arrested the growth of cells harbouring the relevant plasmids.

A similar phenomenon has been described previously by Shimatake and Rosenberg (1981). These workers initially attempted to overproduce the λ*c*II gene in *E.coli* cells under the influence of strong, constitutive promoters, but found that the *c*II gene fragment was invariably inserted in an orientation opposite to the direction of transcription. However, by inserting the gene into a vector carrying the λO_LP_L region, and introducing the recombinant plasmids into *c*I857 lysogens, the lethal function of the *c*II gene was suppressed, and plasmids containing the insert in both orientations were readily obtained. The conclusion drawn from these results was that the amount of repressor synthesised in a *c*I857 lysogen grown under non-inducing conditions was sufficient to reduce *c*II expression to a non-lethal level.

It might have been expected, therefore, that even if the pAD3-4 hybrid insert encoded a harmful phenotype, its expression would be similarly repressed at low temperatures in a *c*I857 host. The present data suggest that perhaps a low level of expression of cloned genes under P_L control does occur in uninduced *c*I857 hosts, and whereas that level of expression may normally be below detection limits, its presence may be detectable in certain circumstances, e.g. if the gene encodes a sufficiently "toxic" product. A similar conclusion was reached by Remaut *et al.* (1983a). To explain the fact that uninduced levels of β-interferon synthesised from λP_L-plasmids was five orders of magnitude lower than in induced cells, whereas uninduced levels of *trp*A were only 300-fold lower than induced levels, these workers advanced the hypothesis that low levels of protein were synthesised under non-inducing conditions in both cases, but that the small amount of the intrinsically unstable β-interferon was degraded to a greater extent by the host cells' proteases.

The deleterious phenotype associated with the pAD3-4 hybrid cDNA was apparently not shared by the pAD2-1 cDNA since after transformation of *c*I857 lysogens with mixtures of recombinant expression plasmids, plasmids containing the latter cDNA in the appropriate orientation for expression were readily isolated. The difference between the two cDNAs which provides the most likely explanation for this phenomenon is the presence of the vicilin signal peptide encoded by pAD3-4. Although it has been demonstrated that certain eukaryotic signal sequences are functional in *E.coli* (Talmadge *et al.*,1980; Fraser and Bruce, 1978), the presence of signal peptides in other proteins merely enables a fraction of these proteins to become attached to the inner membrane (Bassford *et al.*, 1979; Hall and Silhavy, 1981; Kadonaga *et al.*, 1984).

It is possible that the association of a vicilin subunit with the bacterial membrane might, in some way, impair bacterial metabolism. The results of two groups of workers lend support to this contention. Firstly, Remaut *et al.* (1983a) found that several

E.coli strains ceased to grow upon induction of β-interferon synthesis, and suggested that this effect was related to the hydrophobicity of the protein which caused it to stick to the bacterial membrane. Secondly, Hall and Silhavy (1981) have demonstrated that over-production of certain lamB - lacZ, ompF - lacZ and malE-lacZ fusion proteins in *E.coli* causes cell lysis. These fusion products contain the signal peptides of proteins normally exported to the outer membrane or the periplasmic space, and Hall and Silhavy (1981) have provided evidence that lysis was not due to the over-production, *per se*, of the hybrid proteins but was, instead, due to the "jamming" of the bacterial export machinery, and the consequent inability of the cells to efficiently localize large amounts of the hybrid protein. Since it has been shown that strain N99λ*c*I857 lyses under inducing conditions (Fig.23), it is tempting to speculate that the synthesis of vicilin subunits bearing leader sequences renders this strain more prone to cell lysis even under non-inducing conditions.

To further investigate the significance of a leader sequence in relation to the expression of vicilin polypeptides in *E.coli*, the 5' terminus of the insert in pAD3-4.exp1(+) was replaced by the 5' terminus of the pAD2-1 cDNA which contains only two residues of the vicilin signal peptide (see Fig.18). The resulting plasmid, pAD3-4.exp2(+) directed the synthesis, in strain N99λ*c*I857 of a small amount of a ∿47000-*M*r vicilin subunit which was just detectable by Western blotting (Fig.21). This protein was synthesised with an efficiency similar to that of pAD2-1.exp3(+) (N.B. the 5' termini of the vicilin mRNAs produced by pAD2-1.exp3(+) and pAD3-4.exp2(+) are identical), but which was significantly less than that of pAD3-4.exp1(+). This result emphasizes that many factors may potentially affect the yields of proteins synthesised in *E.coli*cells, quite apart from the possible "toxicity" of the encoded product (see section 4.7.3.).

It was subsequently demonstrated that strain K12ΔH1Δ*trp*, transformed with pAD3-4.exp2(+), accumulated approximately twice the levels of the 47000-*M*r vicilin subunits as when transformed with pAD3-4.exp1(+) (see Fig.22). Clearly, several unknown variables are operational in these different plasmid/host strain systems, and it is therefore, difficult to assess from these results, the role played by the vicilin signal peptide in influencing the efficiency of vicilin synthesis.

It is interesting that both pAD3-4.expl(+) and pAD3-4.exp2(+) direct the synthesis of proteins of practically idential *Mr* (∿47000 as estimated by SDS-PAGE,Fig.22) although the former plasmid encodes a vicilin subunit with an extra seventeen amino acids (≡ 2000-*Mr*) at its N-terminus. This suggests the possibility that the vicilin signal peptide may be correctly processed in the bacteria.

4.7.3. Influence of Secondary Structure on Translation of Vicilin mRNAs.

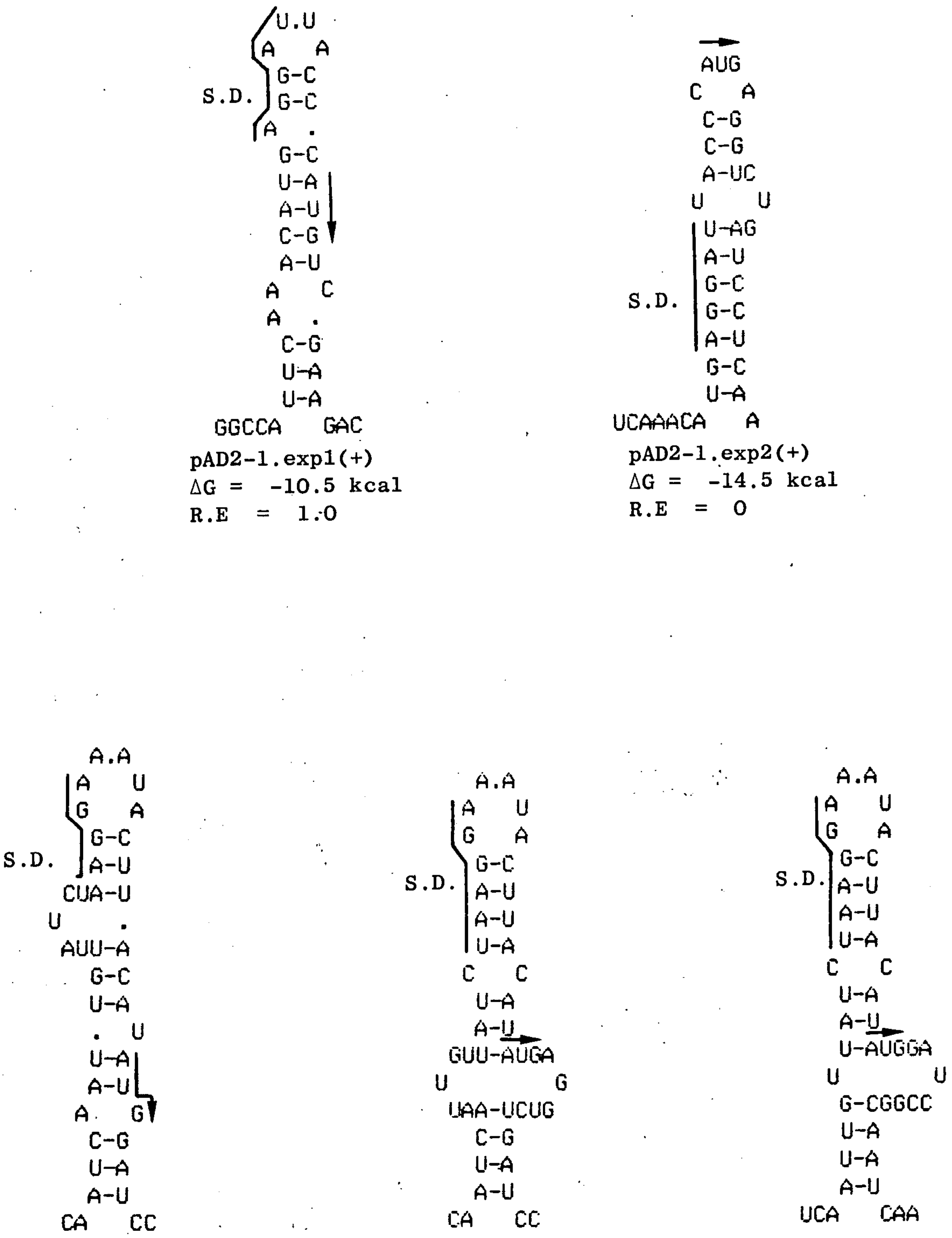
High-level expression of genes in *E.coli* depends on efficient transcription of these genes and on efficient translation of the resulting mRNAs. Since all the expression plasmids constructed in this work were based on the λP_L promoter, the effects of promoter strength in relation to the varying levels of vicilin synthesis obtained can be disregarded. Instead, attention will be focussed on the translational efficiencies of the different mRNAs.

Efficient expression of a bacterial mRNA requires the presence of a "strong" ribosome-binding site (RBS). The major constituents of the RBS are the initiation codon, a purine-rich sequence of 3-9 bases known as the Shine-Dalgarno (SD) sequence, and the spacing (usually 5-10 bases) between the two (Gold *et al.*,1981). The SD sequence is complementary to the 3' end of the 16S rRNA component of the 30S ribosomal subunit (Shine and Dalgarno, 1974) and promotes binding of the bacterial mRNA to the latter during the initiation of translation (Steitz and Jakes, 1975). Despite an overwhelming body of sequence data of known RBS's (see Gold *et al.*, 1981), our understanding of specific factors governing the efficiency of initiation at a given RBS is still very incomplete. However, several studies have suggested that mRNA secondary structures involving the RBS play a major role in determining the efficiency of translation.

To explain the radical variations in the levels of λ*cro* protein synthesised from different *lac-cro* gene fusions, Iserentant and Fiers (1980) examined secondary structure models derived for the various mRNAs and proposed that translational efficiencies were lowered when the RBS, and particularly the initiation codon, was made inaccessible to the ribosomes by sequestration into double-stranded regions of the mRNA. These proposals have been broadly supported by the more recent studies of Wood *et al.* (1984), Tessier *et al.* (1984) and Schottel *et al.* (1984), though the latter workers concluded that it was accessibility (i.e. single-strandedness) of the SD sequence, not the initiation codon, which was of primary importance for efficient translation.

To see whether the differences in the levels of vicilin synthesised by the various plasmid constructs correlated with changes in the likely accessibility of the RBS's, secondary structures were computed using an RNA-folding program (Zuker and Stiegler, 1981; see Fig.28). The vicilin mRNA synthesised by pAD2-1.exp1(+) has the potential to form a fairly stable hairpin structure (ΔG = -10.5 kcal) which buries the initiation codon into its stem but only partially sequesters the SD sequence. By contrast, the hairpin structure derived for the pAD2-1.exp2(+) mRNA (ΔG = -14.5 kcal) sequesters the SD sequence into its stem but exposes the initiation codon within an open loop. The finding that pAD2-1.exp2(+) synthesises very little, if any, vicilin is consistent with the proposal by Schottel *et al.* (1984) that involvement of the SD sequence in base-pairing with neighbouring nucleotides sharply reduces the initiation of translation. To confirm that the lack of expression of vicilin from pAD2-1.exp2(+) is indeed due to the poor translatability of its mRNA, it will be necessary to compare the levels of vicilin-specific mRNA accumulated in cells harbouring the plasmid with cells harbouring, say pAD2-1.exp1(+) (see Schoner *et al.*,1984).

It should be noted that the structure calculated for the pAD2-1.exp(+) mRNA (Fig.28) is characteristic of the initiation region of the phage MS2 replicase gene (Min Jou *et al.*,1972), and not specific for the vicilin sequence. It is known that this particular secondary structure is compatible with efficient translation since the replicase protein is accumulated to high levels in *E.coli* cells when the gene is placed downstream from the λP_L promoter (Remaut *et al.*, 1982). By contrast, the structure derived



pAD2-1.exp1(+)
ΔG = -10.5 kcal
R.E = 1.0

pAD2-1.exp2(+)
ΔG = -14.5 kcal
R.E = 0

pAD2-1.exp3(+)
ΔG = -2.9 kcal
R.E. = 0.2

pAD2-1.exp4(+)
ΔG = -3.8 kcal
R.E. = 0.3

pAD3-4.exp1(+)
ΔG = -3.9
R.E. = 0.2

Figure 28. Computed secondary structures (Zuker and Stiegler, 1981) in the vicinity of the RBS for the vicilin mRNAs synthesised from bacterial plasmids. S.D. sequences are labelled, and the initiation codons are indicated by arrows. R.E. is the relative efficiency of vicilin synthesis. The thermodynamic stability of each structure is as a free energy (ΔG in kcal).

for the pAD2-1.exp2(+) mRNA is specific for the vicilin sequence.

Very similar hairpin structures were calculated for the mRNA molecules synthesised from pAD2-1.exp3(+) (same as pAD3-4.exp2(+)), pAD2-1.exp4(+) and pAD3-4.exp1(+) (see Fig.28) which is consistent with the similarity in vicilin yields produced by all these plasmids. It should be emphasized however, that because of their low ΔG values (-2.9 to -3.9 kcal) it is possible that these theoretical structures do not accurately reflect the structures which actually exist *in vivo*.

It may be concluded that formation of mRNA secondary structures affecting the accessibility of the SD sequence provides a plausible explanation for the differences in vicilin expression seen in cells harbouring the different expression plasmids. However, this conclusion must be treated with caution. For example, the MS2 replicase gene is accumulated to ∿35% of total cell protein when synthesised under the control of the λP_L promoter (Remaut *et al.*, 1982), whereas the MS2 replicase-vicilin fusion product produced by pAD2-1.exp1(+), using the same promoter and translation initiation signals, accumulates to only ∿5%. Thus, it is obvious that other parameters such as the stability of both the mRNA and the translated protein must be involved in determining the efficiency of expression. However, the factors which influence these parameters are not well defined.

REFERENCES

REFERENCES:

1. ALWINE, J.C., KEMP, D.J. and STARK, G.R. (1977). Method for detection of specific RNAs in agarose gels by transfer to diazobenzoyloxy-methyl-paper and hybridisation with DNA probes. Proc. Natl. Acad. Sci. 74, 5350-5354.

2. ARGOS, P., PEDERSEN, K., MARKS, M.D. and LARKINS, B.A. (1982) A structural model for maize zein proteins. J.Biol. Chem. 257, 9984-9990.

3. AUBERT, T.P., BISERTE, G., and LOUCHEUX-LEFEBVRE, M-H.(1976). Carbohydrate-peptide linkage in glycoproteins. Arch. Biochem. Biophys. 175, 410-418.

4. AUBERT, T-P. and LOUCHEUX-LEFEBVRE, M-H. (1976). Conformational study of α_1- acid glycoprotein. Arch. Biochem. Biophys. 175, 400-409.

5. BARTON, K.A., BINNS, A.N., MATCKE, A.J.M. and CHILTON, M-D.(1983). Regeneration of intact tobacco plants containing full length copies of genetically engineered T- DNA, and transmission of T- DNA to R1 progeny. Cell 32, 1033-1043.

6. BARTON, K.A., and BRILL, W.J. (1983). Prospects in plant genetic engineering. Science 219, 671-676.

7. BASSFORD, P.J., SILHAVY, T.J., and BECKWITH, J.R. (1979). Use of gene fusion to study secretion of maltose-binding protein into *Escherichia coli* periplasm. J.Bacteriol. 139, 19-31.

8. BEACHY, R.N., JARVIS, N.P., and BARTON, K.A. (1981). Biosynthesis of subunits of the soybean 7S storage protein. J. Mol. Appl. Genet. 1 , 19-27.

9. BEDBROOK, J.R., SMITH, S.M., and ELLIS, R.J. (1980). Molecular cloning and sequencing of cDNA encoding the precursor to the small subunit of chloroplast ribulose - 1,5-bisphosphate carboxylase. Nature (Lond.) 287, 692-697.

10. BEELEY, J.G. (1977). Peptide chain conformation and the glycosylation of glycoproteins. Biochem. Biophys. Res. Commun. 76, 1051-1055.

11. BERNARD, H-U, REMAUT, E., HERSHFIELD, M.V., DAS, H.K., HELINSKI, D.R., YANOFSKY, C., and FRANKLIN, N. (1979). Construction of plasmid cloning vehicles that promote gene expression from the bacteriophage lambda P_L promoter. Gene 5, 59-76.

12. BEVAN, M.W., FLAVELL, R.B., and CHILTON, M.D. (1983). A chimaeric antibiotic resistance gene as a selectable marker for plant cell transformation. Nature 304, 184-187.

13. BIRNBOIM, H.C., and DOLY, J., (1979). A rapid alkaline extraction procedure for screening recombinant plasmid DNA. Nucl. Acids Res. 7, 1513-1523.

14. BOLIVAR, F., RODRIGUEZ, R.L., GREENE, P.J., BETLOCH, M.V., HEYNECKER, H.L., BOYERm M.W., CROSA, J.H and FALKOW, S., (1977). Construction and characterisation of new cloning vehicles. II. A multipurpose cloning system. Gene 2, 95-113

15. BONNER, W.M. and LASKEY, R.A. (1974). A film detection method for tritium-labelled proteins and nucleic acids in polyacrylamide gels. Eur. J. Biochem. 46, 83-88.

16. BOULTER, D. (1981) Biochemistry of storage protein synthesis and deposition in the developing legume seed. Adv. Bot. Res. 9, 1-31.

17. BRAMMAR, W. J. (1982). Vectors based on bacteriophage lambda. In "Genetic Engineering 3" (Williamson, R., ed.), Academic Press, London. pp.53-81.

18. BRENNAN, C.A., MANTHEY, A.E., and GUMPORT, R.I. (1983). Using T4 RNA ligase with DNA substrates. Methods Enzymol. 100B, 38-52.

19. BRINEGAR, A.C., and PETERSON, D.M. (1982). Synthesis of oat globulin precursors. Analogy to 11S storage protein synthesis. Plant Physiol. 70, 1767-1769.

20. BROGLIE, R., CORUZZI, G., FRALEY, R.T., ROGERS, S.G., HORSCH, R.B., NIEDERMEYER, J.G., FINK, C.L., FLICK, J.S., and CHUA, N-H. (1984). Light-regulated expression of a pea ribulose-1,5-bisphosphate carboxylase small subunit gene in transformed plant cells. Science 224, 838-843.

21. BROWN, J.W.S., ERSLAND, D.R., and HALL, T.C. (1982). Molecular aspects of storage protein synthesis during seed development. In "The Physiology and Biochemistry of Seed Development, Dormancy and Germination" (Khan, A.A., ed.), Elsevier Biomedical Press. pp 3-42.

22. BUCK, K.W., and COUTTS, R.H.A. (1983) Geminiviruses : potential vectors for plant transformation.. Plant Mol. Biol.2, 351-354.

23. BURNETTE, W.N. (1981). "Western blotting" : electrophoretic transfer of proteins from SDS-polyacrylamide gels tó unmodified nitrocellulose and radiographic detection with antibody and radioiodinated protein A. Anal. Biochem. 112, 195-203.

24. CASEY, R., DOMONEY, C., and STANLEY, J. (1984). Convicilin mRNA from pea (*Pisum sativum* L.) has sequence homology with other legume 7S storage protein mRNAs. Biochem J., in press.

25. CASEY, R., MARCH, J.F., and SANGER, E. (1981b). N-terminal amino acid sequence of β-subunits of legumin from *Pisum sativum*. Phytochem 20, 161-163.

26. CASEY, R., MARCH, J.F., SHARMAN, J.E., and SHORT, M.N.,(1981a). The purification, N-terminal amino acid sequence and some other properties of an α^m- -subunit of legumin from the pea (*Pisum sativum* L.) Biochim. Biophys. Acta 670, 428-432.

27. CASEY, R. and SANGER, E. (1980). Purification and some properties of a 7S seed storage protein from *Pisum* (pea). Biochem. Soc. Trans. 8, 657-658.

28. CASEY, R. and SHORT, M.N. (1981). Variation in amino acid composition of legumin from *Pisum*. Phytochem 20, 21-23.

29. CHILTON, M-D., TEPFER, D.A., PETIT, A., DAVIC, C. CASSE-DELBART, F., and TEMPE, J. (1982). *Agrobacterium rhizogenes* inserts T-DNA into the genomes of the host plant root cells. Nature 295, 432-434.

30. CHRISPEELS, M.J. (1984). Biosynthesis, processing and transport of storage proteins and lectins in cotyledons of developing legume seeds. Phil. Trans. R. Soc. Lond. 304B, 309-322.

31. CLEWELL, D.B. (1972). Nature of ColEl plasmid replication in *Escherichia coli* in the presence of chloramphenicol. J. Bacteriol. 110, 667-676.

32. CRAIK, C.S., RUTTER, W.J. and FLETTERICK, R. (1983). Splice junctions : association with variation in protein structure. Science 220, 1125-1129.

33. CROY, R.R.D., DERBYSHIRE I E., KRISHNA, T.G., and BOULTER, D.(1979). Legumin of *Pisum sativum* and *Vicia faba*. New Phytol. 83 29-35.

34. CROY, R.R.D., and GATEHOUSE, J.A. (1985). Genetic engineering of seed proteins—current and potential applications. In "Plant Genetic Engineering," (Dodds, J.H. ed.), Cambridge University Press.

35. CROY, R.R.D., GATEHOUSE, J.A., EVANS, I.M. and BOULTER, D.(1980a). Characterisation of the storage protein subunits synthesised *in vitro* by polyribosomes and RNA from developing pea (*Pisum sativum* L.), 1. legumin. Planta 148, 49-56.

36. CROY. R.R.D., GATEHOUSE, J.A., TYLER, M. and BOULTER, D.(1980b). The purification and characterisation of a third storage protein (convicilin) from the seeds of pea(*Pisum sativum* L.) Biochem. J. 191, 509-516.

37. CROY, R.R.D., LYCETT, G.W., GATEHOUSE, J.A, and BOULTER, D.(1984). Synthesis, cloning and sequence analysis of pea (*Pisum sativum* L.) storage protein specific cDNAs. Kulturpflanze 32, 89-106.

38. CROY, R.R.D., LYCETT, G.W., GATEHOUSE, J.A., YARWOOD, J.N. and BOULTER, D,(1982). Cloning and analysis of cDNAs encoding plant storage protein precursors. Nature (Lond.) 295, 76-79.

39. CROY, R.R.D., SHIRSAT, P. and BOULTER, D., (1985). Cloning and characterisation of genes for legumin from *Pisum sativum* L., in preparation

40. CUNNINGHAM, B.A., WANG, J.L., WAXDOL, M.J. and EDELMAN, G.M.(1975). The covalent and three-dimensional structure of concanavalin A-II amino acid sequence of cyanogen bromide fragment F3 J. Biol. Chem. 250, 1503-1512.

41. DAGERT, M. and EHRLICH, S.D., (1979). Prolonged incubation in calcium chloride improves the competence of *Escherichia coli* cells. Gene 6, 23-28.

42. DANIELSSON, C.E., (1949). Seed globulins of the Gramineae and Leguminosae. Biochem. J..44, 387-400.

43. DARNELL, J.E., (1982). Variety in the level of gene control in eukaryotic cells. Nature 297, 365-371.

44. DAVIES, R.W., (1982), DNA sequencing. In "Gel Electrophoresis of Nucleic Acids - A Practical Approach," (Rickwood, D., and Hames, B.D., eds.), IRL Press Ltd., Oxford. pp.117-172.

45. DE BLOCK, M., HERRERA-ESTRELLA, L., VANMONTAGUE, M., SCHELL, J., and ZAMBRYSKI, P., (1984). Expression of foreign genes in regenerated plants and in their progeny. EMBO J. 3, 1681-1689.

46. DERBYSHIRE, E., WRIGHT, D.J., AND BOULTER, D. (1976). Legumin and vicilin, storage proteins of legume seeds. Phytochem. 15, 3-24.

47. DHAESE, P., DE GREVE, H., GIELEN, J., SEURINCK, J., VAN MONTAGUE, M., AND SCHELL, J., (1983). Identification of sequences involved in the polyadenylation of higher plant nuclear transcripts using *Agrobacterium* T-DNA genes as models. EMBO J. 2, 419-426.

48. DOMONEY, C., and CASEY, R., (1983). Cloning and characterisation of complementary DNA for convicilin, a major seed storage protein in *Pisum Sativum* L. Planta 159, 446-453.

49. DOMONEY, C., AND CASEY, R., (1984). Storage protein precursor polypeptides in cotyledons of *Pisum sativum* L. Identification of, and isolation of a cDNA clone for an 80000-*M*r legumin-related polypeptide. Eur.J.Biochem. 139, 321-327.

50. DRETZEN, G., BELLARD, M., SASSONE-CORSI, P., AND CHAMBON, P. (1981). A reliable method for the recovery of DNA fragments from agarose and acrylamide gels. Anal. Biochem. 112, 295-298

51. EARLY, P., ROGERS. J., DAVIS, M., CALAME, K., BOND, M., WALL., R., and HOOD, L., (1980). Two mRNAs can be produced from a single immunoglobulin μ gene by alternative RNA processing pathways. Cell 20, 313-319.

52. EDENS, L., BOM, I., LEDEBOER, A.M., MAAT, J., TOONEN, M.Y., VISSER, C., and VERRIPS, C.T., (1984). Synthesis and processing of the plant protein thaumatin in yeast. Cell 37, 629-633.

53. EDENS, L., HESLINGA, L., KLOK, R., LEDEBOER, A.M., MAAT, J., TOONEN, M.Y., VISSER, C., AND VERRIPS, C.T. (1982). Cloning of cDNA encoding the sweet plant protein thaumatin and its expression in *Escherichia coli.*, Gene 18, 1-12.

54. EFSTRATIADIS, A., POSAKONY, J.W., MANIATIS, T., LAWN, R.M., O'CONNEL, C., SPRITZ, R.A., DeRIEL, J.K., FORGET, B.G., WEISSMAN, S.M., SLIGHTOM, J.L., BLECHL, A.E., SMITHIES, O., BARALLE, F.E., SHOULDERS, C.C., and PROUDFOOT, N.J (1980). The structure and evolution of the human β-globin gene family. Cell 21, 653-668.

55. ENGLER, M.J., and RICHARDSON, C.C. (1982). DNA ligases. In "The Enzymes" (Boyer, P.D., ed.) 3rd. edition. Academic Press, London. pp.3-29.

56. EVANS, I.M., CROY, R.R.D., BROWN, P. and BOULTER, D. (1980). Synthesis of complementary DNAs to partially purified mRNAs coding for the storage proteins of *Pisum sativum* (L) Biochim. Biophys. Acta 610, 81-95.

57. FAGAN, J.B., PASTAN, I. and deCROMBRUGGHE, B. (1980). Sequence rearrangement and duplication of double stranded fibronectin cDNA probably occuring during cDNA synthesis by AMV reverse transcriptase and *Escherichia coli* DNA polymerase 1 Nucl. Acids. Res. 8, 3055-3064.

58. FEDEROFF, N. (1983). Controlling elements in maize. In "Mobile Genetic Elements." (Shapiro, J., ed). Academic Press, New York. pp.1-63.

59. FITZGERALD, M. and SHENK, T., (1981). The sequence 5'-AAUAAA-3' forms part of the recognition site for polyadenylation of late SV40 mRNAs. Cell 24, 251-260.

60. FLAVELL, R., AND MATHIAS, R.,(1984). Prospects for transforming monocot crop plants. Nature 307, 108-109.

61. FORIERS, A., LEBRUN, E., VAN RAPENBUSCH, R., de NEVE, R., and STROSBERG, A.D., (1981). The structure of the lentil *(Lens culinaris)* lectin-amino acid sequence determination and prediction of the secondary structure. J. Biol. Chem. 256, 5550-5560.

62. FRALEY, R.T., ROGERS, S.G., HORSCH, R.B., SANDERS, P.R., FLICK, J.S., ADAMS, S.P., BITTNER, M.L., BRAND, L.A., FINK, C.L., FRY, J.S., GALLUPPI, G.R., GOLDBERG, S.B., HOFFMAN, N.L., and WOO, S.C, (1983). Expression of bacterial genes in plant cells. Proc. Natl. Acad. Sci. 80, 4803-4807.

63. FRASER, T.H., AND BRUCE, B.J. (1978). Chicken ovalbumin is synthesized and secreted by *Escherichia coli*. Proc. Natl. Acad. Sci. 75, 5936-5940.

64. FORDE, B.G. (1983a). Synthesis of cDNA for molecular cloning. In "Techniques in Molecular Biology," (Walker, J.M. and Gaastra, W., eds.), Croom Helm Publishers, London. pp.167-183.

65. FORDE, B.G. (1983b). Molecular cloning of cDNA : bacterial transformation and screening of transformants. In "Techniques in Molecular Biology," (Walker, J.M. and Gaastra, W. eds.), Croom Helm Publishers, London. pp. 221-238.

66. FYRBERG, E.A., BOND, B.J., HERSHEY, N.D., MIXTER, K.S., and DAVIDSON, N. (1981). The actin genes of *Drosophila* : protein coding regions are highly conserved but intron positions are not. Cell 24, 107-116.

67. GATEHOUSE, J.A., CROY, R.R.D., AND BOULTER, D. (1984). The synthesis and structure of pea storage proteins. CRC Crit. Rev. Plant Sci. 1, 287-314.

68. GATEHOUSE, J.A., CROY, R.R.D., MORTON, H., TYLER, M. and BOULTER, D., (1981). Characterisation and subunit structures of the vicilin storage proteins of pea (*Pisum sativum* L). Eur. J. Biochem. 118, 627-633.

69. GATEHOUSE, J.A., LYCETT, G.W.., CROY, R.R.D., AND BOULTER, D. (1982). The post-translational proteolysis of the subunits of vicilin from pea (*Pisum sativum* L.). Biochem. J. 207, 629-632.

70. GATENBY, A.A., (1983). The expression of eukaryotic genes in bacteria and its application to plant genes. In "Plant biotechnology" (Mantell, S.H. and Smith, H. eds.) Cambridge University Press, Cambridge. pp.269-297.

71. GERAGHTY, D.E., MESSING, J. and RUBENSTEIN, I. (1982). Sequence analysis and comparison of cDNAs of the zein multigene family. EMBO J. 1, 1329-1335.

72. GOLD, L., PRIBNOW, D., SCHNEIDER, T., SHINEDLING, S., SINGER, B.S., and STORMO, G. (1981). Translational initiation in prokaryotes. Ann. Rev. Microbiol. 35 , 365-403.

73. GOODMAN, H.W., and MacDONALD, R.J. (1979). Cloning of hormone genes from a mixture of cDNA molecules. Methods Enzymol. 68, 75-90.

74. GOPINATHAN, K.P., WEYMOUTH, L.A., KUNKEL, T.A., and LOEB, L.A. (1979). Mutagenesis *in vitro* by DNA polymerase from an RNA tumour virus. Nature 278, 857-859.

75. GRUNSTEIN, M., AND WALLIS, J., (1979). Colony hybridisation. Methods Enzymol. 68, 379-389.

76. HALL, M.N. AND SILHAVY, T.L. (1981). Genetic analysis of the major outer membrane proteins of *Escherichia coli*. Ann. Rev. Genet. 15, 91-142.

77. HAMES, B.D. (1981). An introduction to polyacrylamide gel electrophoresis. In "Gel electrophoresis of proteins : a practical approach" (Hames, B.D., and Rickwood, D. eds.) IRL Press LTD., London, pp 1-91.

78. HARRIS, N. and CHRISPEELS, M.J. (1975). Histochemical and biochemical observations on storage protein metabolism and protein body autolysis in cotyledons of germinating mung beans. Plant Physiol. 56, 292-299.

79. HARRIS, T.J.R. (1983). Expression of eukaryotic genes in *E.coli*. In "Genetic Engineering 4" (Williamson R.ed.),Academic Press, London. pp.127-185

80. HARRISON, B. AND ZIMMERMAN, S.B. (1984). Polymer-stimulated ligation : enhanced ligation of oligo- and polynucleotides by T4 RNA ligase in polymer solutions. Nucl. Acids Res. 12, 8235-8251.

81. HEIDECKER, G., and MESSING, J. (1983). Sequence analysis of zein cDNAs obtained by an efficient mRNA cloning method. Nucl. Acids Res. 11, 4891-4906.

82. HELFMAN, D.M., FERAMISCO, J.R., FIDDES, J.C., THOMAS, G.P. and HUGHES, S.H. (1983). Identification of clones that encode chicken tropomyosin by direct immunological screening of a cDNA expression library. Proc. Natl. Acad. Sci. 80, 31-35.

83. HEMPERLY, J.J., HOPP, T.P., BECKER, J.W., and CUNNINGHAM, B.A. (1979). The chemical characterization of favin, a lectin isolated from *Vicia faba*. J. Biol. Chem. 254, 6803-6810.

84. HEMPERLY, J.J., MOSTOV, K.E., and CUNNINGHAM, B.A. (1982). *In vitro* translation and processing of a precursor form of favin, a lectin from *Vicia faba*. J. Biol. Chem. 257, 7903-7909.

85. HERRERA-ESTRELLA, L., DEPICKER, A., VAN MONTAGU, M., and SCHELL, J. (1983a). Expression of chimaeric genes transferred into plant cells using a Ti-plasmid-derived vector. Nature 303, 209-213.

86. HERRERA-ESTRELLA, L., DeBLOCK, M., MESSENS, E., HERNALSTEENS, J.P., VanMONTAGU, M., and SCHELL, J. (1983b). Chimeric genes as dominant selectable markers in plant cells. EMBO J. 2 987-995.

87. HIGGINS, T.J.V. (1984). Synthesis and regulation of major proteins in seeds. Ann. Rev. Plant Physiol. 35, 191-221.

88. HIGGINS, T.J.V., CHANDLER, P.M., ZURAWSKI, G., BUTTON, S.C., and SPENCER, D. (1983a). The biosynthesis and primary structure of pea seed lectin. J. Biol. Chem. 258 , 9544-9549.

89. HIGGINS, T.J.V., CHRISPEELS, M.J., CHANDLER, P.M., and SPENCER, D., (1983b). Intracellular sites of synthesis and processing of lectin in developing pea cotyledons. J. Biol.Chem.258, 9550-9552.

90. HIGGS, D.R., GOODBOURN, S.E.Y., LAMB, J., CLEGG, J.B., WEATHERALL, D.J., AND PROUDFOOT, N.J. (1983). α-thalassaemia caused by a polyadenylation signal mutation. Nature, 306, 398-400.

91. HOFFMAN, L.M., MA,Y. AND BARKER, R.F. (1982). Molecular cloning of *Phaseolus vulgaris* lectin mRNA and use of cDNA as a probe to estimate lectin transcript levels in various tissues. Nucl. Acids. Res. 10, 7819-7828

92. HOPP, T.P., HEMPERLY, J.J. and CUNNINGHAM, B.A. (1982). Amino acid sequence and variant forms of favin, a lectin from *Vicia faba*.J.Biol. Chem. 257, 4473-4483.

93. HORSCH, R.B., FRALEY, R.T., ROGERS., S.G., SANDERS. P.R., LLOYD, A. and HOFFMAN, N. (1984). Inheritance of functional foreign genes in plants. Science 223, 496-498.

94. HU, N-T, PIEFER, M.A., HEIDECKER, G., MESSING, J., and RUBENSTEIN, I., (1982). Primary structure of a genomic zein sequence of maize. EMBO. J. 1, 1337-1342.

95. ISERENTANT, D., and FIERS, W., (1980). Secondary structure of mRNA and efficiency of translation initiation. Gene 9, 1-12.

96. KADONAGA, J.T., GAUTIER, A.E., STRAUS, D.R., CHARLES, A.D., EDGE, M.D., and KNOWLES, T.R.. (1984). The role of the β-lactamase signal sequence in the secretion of proteins by *Escherichia coli* J.Biol. Chem. 259, 2149-2154.

97. KATZ, L., WILLIAMS, P.H., SATO, S., LEAVITT, R.W., and HELINSKI, D.R. (1977). Purification and characterisation of covalently closed replicative intermediates of ColE1 DNA from *Escherichia coli*. Biochemistry 16, 1677-1683.

98. KOUCHALAKOS, R.N., BATES, O.J., BRADSHAW, R.A., and HAPNER, K.D., (1984). Lectin from sainfoin (*Onobrychis viciifolia Scop.*). Complete amino acid sequence. Biochemistry 23, 1824-1830.

99. KROEKER, W.D., KOWALSKI, D., and LASKOWSKI, M. (1976).Mung bean nuclease I. Terminally directed hydrolysis of native DNA. Biochemistry 15, 4463-4467.

100. KÜPPER, H., KELLER, W., KURZ, C., FORSS, S., SCHALLER, H., FRANZE, R., STROHMAIER, K., MARQUARDT, O., ZASLAVSKY, C.G., and HOTSCHNEIDER, P.H. (1981). Cloning of cDNA of major antigen of foot and mouth disease virus and expression in *E.coli*. Nature 289, 555-559.

101. LAEMMLI, U.K. (1970). Cleavage of structural proteins during the assembly of the head of bacteriophage T4. Nature 227, 680-685.

102. LAMB, I. (1984). The cloning and characterisation of cDNAs encoding castor bean lectins. Ph.D. thesis, University of Warwick, UK.

103. LAND, H., GREZ, M., HAUSER, H., LINDENMAIER, W., and SCHÜTZ, G., (1981). 5'-terminal sequences of eukaryotic mRNA can be cloned with high efficiency. Nucl. Acids Res. 9, 2251-2266.

104. LARKINS, B.A. (1981). Seed storage proteins : characterization and biosynthesis. In "The Biochemistry of Plants - A Comprehensive Treatise," (Stumpf. P.K. and Conn, E.E., eds.),"Vol 6, Proteins and Nucleic Acids," (Marcus, A. ed.), Academic Press, London. pp.449-489.

105. LARKINS, B.A. (1983). Genetic engineering of seed storage proteins. In "Genetic Engineering of Plants," (Hollaender, A., Kosuge, T. and Meredith, C.P. eds.). Plenum Press, New York, 93-118.

106. LASKEY, R.A., and MILLS, A.D., (1975). Quantitative film detection of ^{3}H and ^{14}C in polyacrylamide gels by fluorography. Eur J. Biochem. 56, 335-341.

107. LATHE, R.F., LECOCQ, J.P., and EVERETT, R. (1983). DNA engineering : the use of enzymes, chemicals and oligonucleotides to restructure DNA sequences *in vitro.* In "Genetic Engineering 4" (Williamson, R. ed.). Academic Press, London. pp.1-56.

108. LEEMANS, J., SHAW, C.H., DEBLAERE, R., DeGREVE, H., HERNALSTEENS, J.P., MAES, M., VanMONTAGU, M. and SCHELL, J. (1981). Site-specific mugagenesis of *Agrobacterium* Ti plasmids and transfer of genes to plant cells. J.Mol.Appl. Genet.1, 149-164.

109. LYCETT, G.W., CROY, R.R.D., SHIRSAT, A.H. and BOULTER, D. (1984a). The complete nucleotide sequence of a legumin gene from pea(*Pisum sativum* L.) Nucl. Acids Res. 12, 4493-4506.

110. LYCETT, G.W., DELAUNEY, A.J., and CROY, R.R.D., (1983b). Are plant genes different? FEBS lettr. 153, 43-46.

111. LYCETT, G.W., DELAUNEY, A.J., GATEHOUSE, J.A., GILROY, J., CROY, R.R.D., and BOULTER, D. (1983a). The vicilin gene family of pea (*Pisum sativum* L.) : a complete cDNA coding sequence for preprovicilin. Nucl. Acids.Res. II, 2367-2380.

112. LYCETT, G.W., DELAUNEY, A.J., ZHAO, W., GATEHOUSE, J.A., CROY, R.R.D., and BOULTER, D. (1984b). Two cDNA clones coding for the legumin protein of *Pisum sativum* L. contain sequence repeats. Plant Mol. Biol. 3, 91-96.

113. MAAT, J., and SMITH, A.J.H. (1978). A method for sequencing restriction fragments with dideoxynucleoside triphosphates. Nucl. Acids Res. 5, 4537-4546.

114. MANIATIS, T., FRITSCH, E.F., and SAMBROOK, J. (1982). Molecular Cloning - A Laboratory Manual (Cold Spring Harbour Laboratory, New York). 545pp.

115. MARSHALL, R-D. (1972). Glycoproteins. Ann. Rev. Biochem. 41, 673-702.

116. MATTA, N.K., GATEHOUSE, J.A., AND BOULTER, D. (1981). Molecular and subunit heterogeneity of legumin of *Pisum Sativum* L. (garden pea)- a multi-dimensional gel electrophoretic study. J. Exp. Bot. 32, 1295-1307.

117. MATZKE, A.J.M. and CHILTON, M-D. (1981). Site-specific insertion of genes into T-DNA of the *Agrobacterium* tumor-inducing plasmid : an approach to genetic engineering of higher plant cells. J. Mol. Appl. Genet. 1, 39-49.

118. MAXAM, A.M., AND GILBERT, W. (1980). Sequencing end-labelled DNA with base-specific chemical cleavages. Methods Enzymol. 65, 499-560.

119. MESSING, J. (1983). New M13 vectors for cloning. Methods Enzymol. 101C, 20-78.

120. MESSING, J., GERAGHTY, D., HEIDECKER, G., HU, N-T., KRIDL, J., and RUBENSTEIN, I. (1983). Plant gene structure. In "Genetic Engineering of Plants", (Hollaender, A., Kosuge, T. and Meredith, C.P. eds). Plenum Press, New York, pp. 211-227.

121. MIN JOU, W., HAEGEMAN, G., TSEBAERT, M. and FIERS, W. (1972). Nucleotide sequence of the gene coding for the bacteriophage MS2 coat protein. Nature 237, 82-88.

122. MONTELL, C., FISHER, E.F., CARUTHERS, M.H., AND BERK, A.J. (1983). Inhibition of RNA cleavage but not polyadenylation by a point mutation in mRNA 3' consensus sequence AAUAAA Nature 305, 600-605.

123. MORTON, H., EVANS, I.M., GATEHOUSE, J.A., AND BOULTER, D. (1983). Sequence complexity of messenger RNA in cotyledons of developing pea (*Pisum sativum*) seeds. Phytochem. 22, 807-812.

124. MURAI, N., SUTTON, D.W.., MURRAY, M.G., SLIGHTOM, J.L., MERLO, D.J., REICHERT, N.A., SINGUPTA-GOPALAN, C., STOCK, C.A., BARKER, R.F., KEMP, T.D., AND HALL, T.C.. (1983). Phaseolin gene from bean is expressed after transfer to sunflower via tumour-inducing plasmid vectors. Science 222, 476-482.

125. NELSON, T., AND BRUTLAG, D. (1979). Addition of homopolymers to the 3'-ends of duplex DNA with terminal transferase. Methods Enzymol. 68, 41-51.

126. NIELSEN, N.C. (1984). The chemistry of legume storage proteins. Phil. Trans. R. Soc. Lond. 304 B, 287-296.

127. OKAYAMA, H. and BERG, P. (1982). High-efficiency cloning of full-length cDNA. Mol. Cell. Biol., 2, 161-170.

128. O'MALLEY, B.W., TOWLE, H.C. and SCHWARTZ, R.J., (1977). Regulation of gene expression in eukaryotes. Ann. Rev. Genet. 11, 239-275.

129. ORAM, R.N. and BROCK, R.D., (1972). Prospects for improving plant protein yield and quality by breeding. J.Aust. Inst. Agric. Sci. 38, 163-168.

130. OSBORNE, T.B. (1924). The Vegetable Proteins. 2nd. ed. (Longmans. Green and Co., London.). 154 pp.

131. PROUDFOOT, N.J. (1982). The end of the message. Nature 298, 516-517.

132. RASMUSSEN, S.K., HOPP, H.E., and BRANDT, A. (1983). Nucleotide sequences of cDNA clones for B1 hordein polypeptides. Carlsberg Res. Commun. 48, 187-199.

133. REMAUT, E., DeWAELE, P., MARMENOUT, A., STANSSENS, P., and FIERS, W. (1982). Functional expression of individual plasmid-coded RNA bacteriophage MS2 genes. EMBO. J.1, 205-209.

134. REMAUT, E., STANSSENS, P., AND FIERS, W. (1981). Plasmid vectors for high-efficiency expression controlled by the P_L promoter of coliphage lambda. Gene 15, 81-93.

135. REMAUT, E., STANSSENS, P. and FIERS, W. (1983a). Inducible high level synthesis of mature human fibroblast interferon in *Escherichia coli*. Nucl. Acids Res. 11, 4677-4688.

136. REMAUT, E., TSAO, H. and FIERS, W. (1983b). Improved plasmid vectors with a thermoinducible expression and temperature-regulated runaway replication. Gene 11, 103-113.

137. RICHARDSON, M., CAMPOS, F.D.A.P., MOREIRA, R.A., AINOUZ, I.L., BEGBIE, R., WATT, W.B., and PUSZTAI, A. (1984). The complete amino acid sequence of the major α subunit of the lectin from the seeds of *Dioclea grandiflora* (Mart). Eur. J. Biochem. in press.

138. RIGBY, P.W.J., DIECKMANN, M., RHODES, C., and BERG, P. (1977). Labelling deoxyribonucleic acid to high specific activity *in vitro* by nick translation with DNA polymerase 1. J. Mol. Biol. 113, 237-251.

139. ROSENBERG, M., HO, Y-S., and SHATZMAN, A. (1983). The use of pKC30 and its derivatives for controlled expression of genes. Methods Enzymol. 101 , 123-138.

140. ROUGEON, F., KOURILSKY, P. and MACH, B. (1975). Insertion of a rabbit β-globin gene sequence into an *E.coli* plasmid. Nucl. Acids Res. 2, 2365-2378.

141. RUBIN, G.M. and SPRADLING, A.C. (1982). Genetic transformation of *Drosophila* with transposable element vectors. Science 218, 348-353.

142. SCHOLZ, G., MANTEUFFEL, R., MUNTZ, K. and RUDOLPH, A. (1983). Low-molecular-weight polypeptides of vicilin from *Vicia faba* L. are products of proteolytic breakdown. Eur. J. Biochem. 132, 103-107.

143. SCHONER, B.E., HSLUING, H.M. HELAGAJE, R.M., MAYNE, N.G. and SCHONER, R.G.. (1984). Role of mRNA translational efficiency in bovine growth hormone expression in *Escherichia coli*. Proc. Natl. Acad. Sci. 81, 5403-5407.

144. SCHOTTEL, J.L., SNINSKY, J.L. and COHEN, S.N. (1984). Effects of alterations in the translation control region on bacterial gene expression : use of *cat* gene constructs transcribed from the *lac* promoter as a model system. Gene 28, 177-193.

145. SCHROEDER, H.E. (1982). Quantitative studies on the cotyledonary proteins in the genus *Pisum* J.Sci. Food Agric. 33, 623-633.

146. SCHULER, M.A., LADIM, B.F., POLLACO, J.C. FREYER, G. and BEACHY, R.N. (1982a). Structural sequences are conserved in the genes coding for the α, α' and β-subunits of the soybean 7S seed storage protein. Nucl. Acids Res. 10, 8245-8261.

147. SCHULER, M.A., SCHMITT, E.S., and BEACHY, R.N. (1982b). Closely related families of genes code for the α and α' subunits of the soybean 7S storage protein complex. Nucl. Acids Res. 10, 8225-8244.

148. SEIF, I., KHOURY, G. and DHAR, R. (1980). A rapid enzymatic DNA sequencing technique : determination of sequence alterations in early simian virus 40 temperature sensitive and deletion mutants. Nucl. Acids Res.8, 2225-2239.

149. SETZER, D.R., McGROGAN, M., and SCHIMKE, R.T. (1982). Nucleotide sequence surrounding multiple polyadenylation sites in the mouse dihydrofolate reductase gene. J. Biol. Chem. 257, 5143-5147.

150. SHAW, C.H. (1984). Ti-plasmid-derived plant gene vectors. Oxf. Surveys Plant Mol. Cell Biol. 1, 211-216.

151. SHAW, C.H., LEEMANS, J., SHAW, C.H., van MONTAGU, M. and SCHELL, J. (1983). A general method for the transfer of cloned genes to plant cells. Gene 23, 315-330.

152. SHENK, T.E., RHODES, C., RIGBY, P.W. AND BERG, P. (1975). Biochemical method for mapping mutational alterations in DNA with S1 nuclease : the location of deletions and temperature-sensitive mutations in simian virus 40. Proc. Natl. Acad. Sci. 72, 989-993.

153. SHEWRY, P.R., MIFLIN, B.J., FORDE, B.G., and BRIGHT, S.W.J. (1981). Conventional and novel approaches to the improvement of the nutritional quality of cereal and legume seeds. Sci. Prog. Oxf. 67, 575-600.

154. SHIMATAKE, H. and ROSENBERG, M. (1981). Purified λ regulatory protein *c*II positively activates promoters for lysogenic development. Nature 292, 128-132.

155. SHINE, J. and DALGARNO, L. (1974). The 3'-terminal sequence of *Escherichia coli* 16S ribosomal RNA : complementarity to nonsense triplets and ribosome binding sites. Proc. Natl Acad. Sci. 71, 1342-1346.

156. SHIRSAT, A.H. (1984). A gene for legumin : a major storage protein of *Pisum Sativum* L. Ph.D. thesis, University of Durham.

157. SIPPEL, A.E., LAND. H., LINDENMAIER, W., NGYYEN-HUU, M.C., WURTZ, T., TIMMIS, K.N., GIESECKE, K. and SCHUTZ, G. (1978). Cloning of chicken lysozyme structural gene sequences synthesised *in vitro*. Nucl. Acids Res. 5,3275-3294.

158. SLIGHTOM, J.L., SUN, S.M. and HALL, T.C. (1983). Complete nucleotide sequence of a French bean storage protein gene : phaseolin. Proc. Natl. Acad. Sci. 80, 1897-1901.

159. SMITH, D.F., SEARLE, P.F. and WILLIAMS, J.G. (1979). Characterisation of bacterial clones containing DNA derived from vitellogenin mRNA. Nucl. Acids Res. 6, 487-506.

160. SORENSON, J.C. (1984). The structure and expression of nuclear genes in higher plants. Adv. Genet. 22, 109-144.

161. SOROKIN, A.F., PETRENKO, O.I., KAVSAN, V.M. KOXLOV, Y.I., DEBABOV, V.G. and ZLOCHEVSKIJ, M.L. (1982). Nucleotide sequence analysis of the cloned salmon preproinsulin cDNA. Gene 20, 367-376.

162. SOUTHERN, E.M. (1975). Detection of specific sequences among DNA fragments separated by gel electrophoresis. J.Mol. Biol. 98, 503-517.

163. SPENCER, D., CHANDLER, P.M. HIGGINS, T.J.V., INGLIS, A.S., and RUBIRA, M. (1983). Sequence interrelationships of the subunits of vicilin from pea seeds. Plant Mol. Biol., 2, 259-267.

164. SPENCER, D., and HIGGINS, T.J.V. (1980). The biosynthesis of legumin in maturing pea seeds. Biochem. Int. 1, 502-509.

165. STEITZ, J.A, and JAKES, K. (1975). How ribosomes select initiator regions in mRNA : Base pair formation between the 3' terminus of 16S rRNA and mRNA during initiation of protein synthesis in *Escherichia coli*. Proc. Natl. Acad. Sci. 72, 4734-4738.

166. STRUCK, D.K., LENNARZ, W.J. and BREW, K. (1978). Primary structural requirements for the enzymatic formation of the N-glycosidic bond in glycoproteins. J. Biol. Chem. 253, 5786-5794.

167. STUDIER, F.W. (1973). Analysis of bacteriophage T7 Early RNAs and proteins on slab gels. J. Mol. Biol. 79, 237-248.

168. SUGINO, A., GOODMAN, H.M., HEYNEKER, H.L., SHINE, J., BOYER, H.W. and COZZARELLI, N.R. (1977a). Interaction of bacteriophage T4 RNA and DNA ligases in joining of duplex DNA at base-paired ends. J. Biol. Chem. 252, 3987-3994.

169. SUGINO, A., SNOPEK, T.J. and COZZARELLI, N.R. (1977b). Bacteriophage T4 RNA ligase - reaction intermediates and interaction of substrates. J. Biol. Chem. 252, 1732-1738.

170. Sutcliffe, J.G. (1978). Complete nucleotide sequence of the *Escherichia coli* plasmid pBR322. Cold Spring Harb. Symp. Quant. Biol. 43, 77-90.

171. TALMADGE, K., KAUFMAN, J. and GILBERT, W. (1980). Bacteria mature preproinsulin to proinsulin. Proc. Natl. Acad. Sci. 77, 3988-3992.

172. TESSIER, L-H. SONDERMEYER, P., FAURE, T., DREYER, D., BENAVENTE, A., VILLEVAL, D., COURTNEY, M. and LECOCQ, J.P. (1984). The influence of mRNA primary and secondary structure on human IFN-γ gene expression in *E.coli*. Nucl. Acids Res. 12, 7663-7675.

173. TOSI, M., YOUNG, R.A., HAGENBUCHLE, O. and SCHIBER, U.(1981). Multiple polyadenylation sites in a mouse α-amylase gene. Nucl. Acids Res. 9, 2313-2323.

174. TOWBIN, H., STAEHELIN, T. and GORDON, J. (1979). Electrophoretic transfer of proteins from polyacrylamide gels to nitrocellulose sheets : procedure and some applications. Proc. Natl. Acad. Sci. 76, 4350-4354.

175. ULMER, K.M. (1983). Protein engineering. Science 219, 666-671.

176. VanEMBDEN, J. (1983). The use of cosmids as cloning vehicles. In "Techniques in Molecular Biology" (Walker, J.M. and Gäastra, W. eds.), Croom Helm Publishers, London, pp.309-321.

177. VILLA-KOMAROFF, L., EFSTRATIADIS, A., BROOME, S., LOMEDICO, P., TIZARD, R., NABER, S.P., CHICK, W.L. and GILBERT, W. (1978). A bacterial clone synthesizing proinsulin. Proc. Natl. Acad. Sci. 75, 3727-3731.

178. 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.

179. VOLCKAERT, G., TAVERNIER, J., DERYNCK, R., DEVOS, R., and FIERS, W. (1981). Molecular mechanisms of nucleotide-sequence rearrangements in cDNA clones of human fibroblast interferon mRNA. Gene 15, 215-223.

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

181. WANG J.L., CUNNINGHAM, B.A., WAXDÄL, M.J. and EDELMAN, G.M. (1975). The covalent and three-dimensional structure of concanävalin A - I-amino acid sequence of cyanogen bromide fragments Fl and F2. J. Biol. Chem. 250, 1490-1502.

182. WEAVER, C.A. GORDON, D.F. and KEMPER, B. (1981). Introduction by molecular cloning of artifactual inverted sequences at the 5' terminus of the sense strand of bovine parathyroid hormone cDNA. Proc. Natl. Acad. Sci. 78, 4073-4077.

183. WENDORF, F., SCHILD, R., EL HADIDI, N., CLOSE, A.E., KOBUSIEWICZ, M., WIECKOWSKA, H., ISSAWI, B. and HAAS, H. (1979). Use of barley in the Egyptian late paleolithic. Science 205, 1341-1347.

184. WEN-MING, Z., GATEHOUSE, J.A. and BOULTER, D. (1983). The purification and partial amino acid sequence of a polypeptide from the glutelin fraction of rice grains ; homology to pea legumin. FEBS Lettr. 162, 96-102.

185. WICKENS, M.P., BUELL, G.N. and SCHIMKE -R.T. (1978). Synthesis of double-stranded DNA complementary to lysozyme, ovomucoid, and ovalbumin mRNAs - Optimisation for full length second strand synthesis by *Escherichia coli* DNA polymerase 1 J. Biol. Chem. 253, 2483-2495.

186. WILLIAMS, J.G. (1981). The preparation and screening of a cDNA clone bank. In "Genetic Engineering 1" (Williamson, R. ed.) Academic Press, London. pp.1-59.

187. WOOD, C.R. BOSS, M.A., PATEL, T.P. and EMTAGE, T.S.,(1984). The influence of messenger RNA secondary structure on expression of an immunoglobulin heavy chain in *Escherichia coli*. Nucl. Acids Res. 12, 3937-3950.

188. WRIGHT, D.J. and BOULTER, D. (1974). Purification and subunit structure of legumin of *Vicia faba* (broad bean). Biochem. J. 141, 413-418.

189. YAMAGATA, H., SUGIMOTO, T., TANAKA, K. and KASAI, Z.(1982). Biosynthesis of storage proteins in developing rice seeds. Plant Physiol. 70, 1094-1100.

190. YOUNG, J.F., DESSELBERGER, U., PALESE, P., FERGUSON, B., SHATZMAN, A.R. and ROSENBERG, M. (1983). Efficient expression of influenza virus NS1 nonstructural proteins in *Escherichia coli*. Proc. Natl. Acad. Sci. 80, 6105-6109.

191. ZUKER., M. and STIEGLER, P. (1981). Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucl. Acids Res. 9, 133-148.