Volume 13 Number 18 1985 Nucleic Acids Research

The 5'-flanking regions of three pea legumin genes: comparison of the DNA sequences

Grantley W.Lycett[1*], Ronald R.D.Croy, Anil H.Shirsat, D.Margaret Richards and Donald Boulter

Department of Botany, University of Durham, South Road, Durham DH1 3LE, UK

Received 17 June 1985; Revised and Accepted 3 September 1985

ABSTRACT

Approximately 1200 nucleotides of sequence data from the promoter and 5'-flanking regions of each of three pea (Pisum sativum L.) legumin genes (legA, legB and legC) are presented. The promoter regions of all three genes were found to be identical including the 'TATA box', and 'CAAT box', and sequences showing homology to the SV40 enhancers. The legA sequence begins to diverge from the others about 300bp from the start codon, whereas the other two genes remain identical for another 550bp. The regions of partial homology exhibit deletions or insertions and some short, comparatively well conserved sequences. The significance of these features is discussed in terms of evolutionary mechanisms and their possible functional roles. The legC gene contains a region that may potentially form either of two mutually exclusive stem-loop structures, one of which has a stem 42bp long, which suggests that it could be fairly stable. We suggest that a mechanism of switching between such alternative structures may play some role in gene control or may represent the insertion of a transposable element.

INTRODUCTION

The legumin protein of Pisum sativum L., one of the two major pea seed storage proteins, is synthesised exclusively in the developing embryo, particularly in the cotyledons, during the period 8 days to 20 days after flowering. The synthesis of the protein, which closely follows the level of mRNA, is almost certainly regulated at the transcriptional level (1, 2, 3, 4). The legumin gene is therefore a promising system for the study of developmental regulation of gene expression.

One member of the pea legumin gene family has recently been isolated and sequenced (5). The gene sequence was shown to be identical to that of the legumin cDNA pDUB8 and the encoded amino acid sequence matched extensive data from the major legumin polypeptides, suggesting that the gene is functional in vivo (5, 6). Additionally the same gene has been shown to be active in the Hela cell in vitro transcription system (7). The sequence requirements for tissue-specific legumin gene expression are not known though the 'TATA box' and 'CAAT box' characteristic of animal genes have been found in the legumin gene and transcription has been shown to start 24bp downstream of the 'TATA box' (5, 7). In the Agrobacterium tumefaciens Ti plasmid system it has been shown that both the 'CAAT' and 'TATA' boxes are required for transcriptional activity of the nopaline synthase gene in plants (8). No control sequences have yet been identified in plant genes though the requirements for regulated and tissue specific expression of the ribulose bisphosphate carboxylase (RUBISCO) gene appear to reside somewhere on a 2.3 kb fragment of DNA containing the gene (9). The enhancer elements found in some animal genes and animal virus genes appear to show tissue specificity (10) and similar sequences may play a role in the control of plant gene expression. A region with homology to the core of the adenovirus enhancer has been found in the pea legA gene (5) and homology to the SV40 enhancer core has been shown in the RUBISCO gene (9) but many other genes do not show such homologies and no functional significance has been demonstrated for such sequences in a plant system.

In the belief that any sequences involved in the control of gene expression are likely to show evolutionary conservation within groups of similarly regulated genes, we have sequenced the promoter regions of two other pea legumin genes (legB and legC) for comparison with that of the legA gene (5). We have also determined the sequences extending more that 1kbp upstream of the promoters of all three genes. Comparison of the three genes showed all three promoter regions to be identical, but more variable regions surrounding some short conserved sequences were identified further upstream. The 5' end of the legC sequence was found to contain a complex region of direct and inverted repeats which has the potential to form either of two mutually exclusive stem-loop structures.

MATERIALS AND METHODS

Materials

Sources of materials were as listed previously (5). Genomic DNA was isolated according to the method of Graham (11) from leaf tissue of 10-11 day old pea seedlings (Pisum sativum L. cv Feltham First) and purified twice by centrifugation in caesium chloride-ethidium bromide density gradients.

Genomic clones Genomic DNA was subjected to restriction by either EcoR I or Sau3A I under conditions where partial cleavage occurred, as judged by gel electrophoretic analyses (12). Restricted DNAs were size fractionated on 10 - 40% (w/v) sucrose density gradients, the fractions analysed by gel electrophoresis and the appropriate range of fragment sizes cloned in λ L47.1 (13) for Sau3A I fragments or λgt WES. λB (14) for EcoR I fragments. Legumin genomic clones were isolated from the libraries using the 1100bp legumin cDNA excised from pDUB6 with BamH I for screening (6). Three genomic clones were selected for restriction mapping (designated λ leg 1, 2 and 3) and the legumin gene sequences located and orientated on southern blots using cDNA fragments specific for the 5' and 3' ends of the coding sequence (15). Restriction fragments containing the complete coding, 5' and 3' flanking sequences of all three genes were subcloned into pUC8 (16) as follows: legA on a 3.3 kbp, BamH I fragment from λ leg 1 clones as pDUB24; legB on a 3.8 kbp EcoR I- BamH I fragment from λ leg 2 as pDUB25; legC on a 4.7 kbp EcoR I - BamH I fragment from λ leg 3 as pDUB27. Detailed restriction mapping of these fragments (12) directed further subcloning into pUC8 or M13 mp8 (17) for sequencing.

DNA sequencing All three DNA clones were sequenced completely by the 'forwards and backwards' dideoxy nick translation method of Seif *et al.*, (18) with modifications described in detail previously (5). Two short sequences

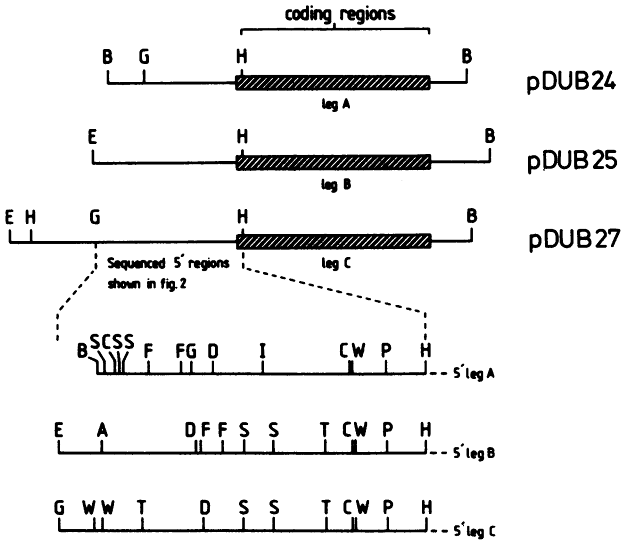

Figure 1 Fragment positions and detailed restriction maps of the 5' flanking sequences of the legumin genes in clones λ leg 1, 2 and 3 and in subclones pDUB24, 25 and 27. The symbols used for restriction enzyme recognition sites are: A = Acc I; B = BamH I; C = Nco I; D = Nde I; E = EcoR I; F = Hinf I; G = Bgl II; H = Hind III; I = Hpa II; P = Pst I; S = Sau3A I; T = BstE II; W = Ava II.

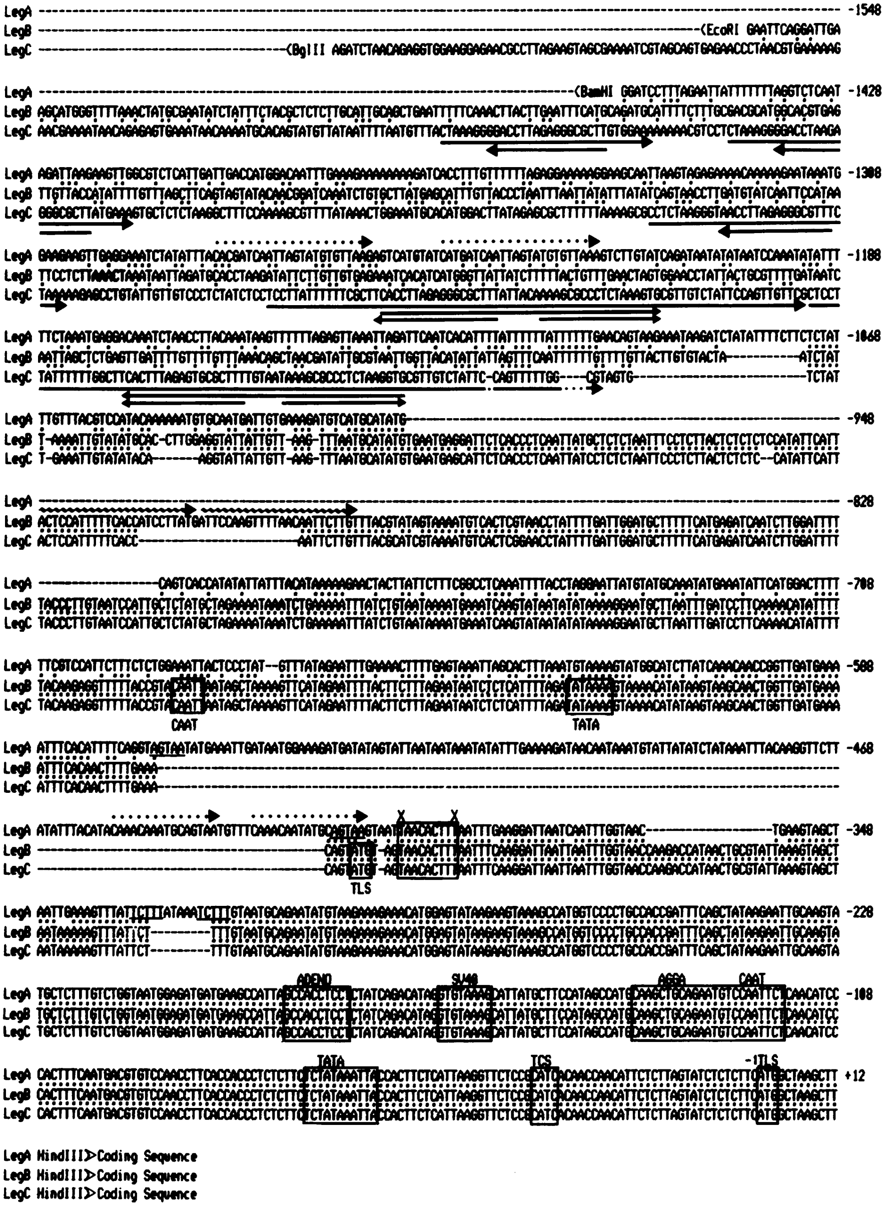

Figure 2 The complete nucleotide sequences of the 5' flanking regions of legumin genes A, B and C with homologous sequences aligned by visual inspection and according to dot matrix comparisons as shown in figure 3. Identical bases in the different gene sequences are denoted by dots. Deletions are shown as broken lines within sequences. Direct and inverted sequence repeats are indicated by arrowed lines for legA (·················▶),

legB (⟿) and *legC* (⟶). Putative control and start sequences are enclosed in boxes. Abbreviations used are TLS - translation start; TCS = transcription start; SV40 = sequence homologous to an SV40 enhancer; ADENO = sequence homologous to an adenovirus enhancer; X X = sequence homologous to a soybean seed protein 5' gene sequence (Goldberg and Sims, personal communication); TATA, AGGA, CAAT = consensus sequences as found in other eukaryotic gene sequences. Other features of potential significance are underlined.

were confirmed by dideoxy-termination sequencing (19) after cloning into M13 mp8 (17).

RESULTS AND DISCUSSION

The 5' non-transcribed regions of three pea legumin genes, designated *legA*, *legB* and *legC*, have been sequenced and compared. Restriction maps of the relevant regions are shown in figure 1. The production and characterisation of the lambda genomic clones and the pUC8 plasmid sub-clones is described in more detail elsewhere (15) (Croy *et al*., 1985, manuscript in preparation). The sequence of the coding region of the *legA* gene, together with 200 nucleotides upstream from the translation start have been presented previously (5). The insert in plasmid subclone pDUB24 has now been completely sequenced to provide an additional 1000 nucleotides of 5' flanking sequence and the corresponding 5' flanking regions and promoter regions of two other legumin genes *legB* and *legC*, have also been sequenced for comparison. The complete sequences are shown in figure 2 where they have been aligned by the introduction of gaps to maximize homology. The alignment was aided by dot matrix comparisons (figures 3A and 3B) performed to provide a quantitative assessment of the extent and degree of homology between the three sequences. The present paper shows the sequence of *legA*, corrected at position -87 for a single base error copied in the sequence presented in (5).

The 321bp of sequence proximal to the coding regions of all three genes were found to be identical. This part of the *legA* gene has already been shown to contain 'TATA' and 'AGGA/CAAT' boxes and also a sequence homologous to an adenovirus enhancer (5). Further sequence searching has revealed the presence of the sequence GTGTAAAG (position -160 to -167) which is 90% homologous to the SV40 enhancer core sequence reported by Weiher *et al*. (20). Thus these features are common to all three genes. Although it might have been expected that functional, structural features would be preferentially conserved whilst non-functional regions diverged during evolution, this was not the type of pattern observed in the *leg* gene promoter regions and therefore no additional

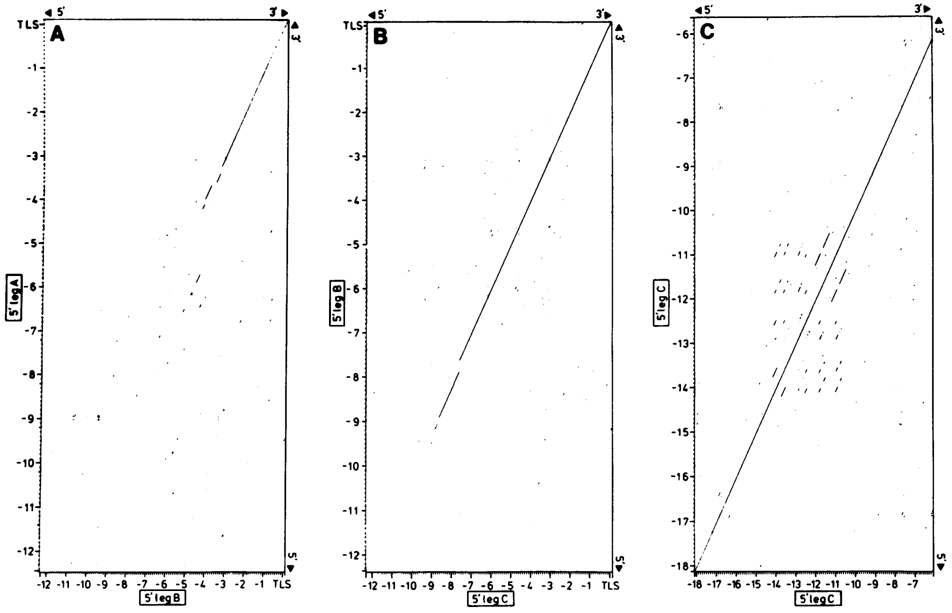

Figure 3 Dot matrix comparisons of the three nucleotide sequences of the 5' flanking regions of: A) legA with legB and B) legB with legC. The vertical and horizontal scales are in nucleotides (x 10^{-2}) upstream from the translation start (TLS). The comparison of legA with legC was closely similar to that between legA and legB (fig. 3A) due to the extensive homology of the legB and legC flanking sequences (data not presented). C) is a comparison of the legC sequence from position -500 to -1800, with itself (self-comparison) illustrating the extensive region of repeated sequences. Comparisons were computed on a BBC model B microcomputer using a programme based on that of Staden (32). Each 'dot' in the matrix represents a true match of at least eight bases in a sequence scan (window) of ten bases.

support may be lent to observations on the functional significance of features located in the promoter region (5; figure 2). The question of the expression of the legumin genes still remains. The evidence for expression of the legA gene in vivo and in vitro has been presented elsewhere (5, 7) and supports the contention that this gene is active to a high level in the developing seed. The evidence for the expression of the legB and legC genes is somewhat more circumstantial. The major legumin protein consists in most genetic lines of peas, of three or four species of polypeptides differing only slightly in size and composition (21). The legumin gene family thought to encode these polypeptides is made up of only a small number (4-6 members) of genes (22, 23). It is probable therefore that several of the genes in addition to legA are functional. Furthermore the high degree of homology between the 5' flanking, (present paper) the coding and 3' flanking sequences (J. Gatehouse, personal communication) with the apparently expressed *legA* gene tends to support the liklihood of their expression.

As may be seen from figure 3, the gene sequences do show some divergence upstream of position -322, although extensive regions of homology, some of which may be functional, are still present. Between positions -322 and -700 the *legA* sequence begins to diverge from *legB* and *legC*. A long stretch (146bp) of extra sequence is present in *legA* between positions -424 and -570 representing either a major insertion in *legA* or a deletion from *legB* and *legC*. One of the last regions that shows good homology between all three genes may be seen between positions -377 and -424. This regions contains the sequence TAACACTTT (-405 to -413) which closely resembles a consensus sequence of the form (A/T/C)AACACA(AA/CT) located in several soybean seed specific genes including glycinin (soybean legumin), lectin, β-conglycinin and trypsin inhibitor genes (R. Goldberg and T. Sims, personal communication). Immediately upstream from the position of the inserted/deleted sequence are some well conserved sequences, for example between -580 and -606 and between -659 and -686. It is also interesting to note that a second set of promoter-like sequences occurs in this region in the *legB* and *legC* sequences with a 'CAAT box' (CAATT) at -687 and a 'TATA box' (TATAAAA) at -628. These are followed by an appropriately positioned ATG, (translation start) triplet, at -420 (figure 2) but this is in turn closely followed by several TAA (translation stop) triplets in the same reading frame precluding transcription of a functional message from such promoters.

Beyond position -700 the *legA* sequence diverges completely from the *legB* and *legC* sequences while these two remain identical up to position -874. Only upstream from this position do they begin to show divergence. It is of interest to note that some sequences between positions -1013 and -1122 in one of the last regions of good homology between *legB* and *legC*, can be aligned with sequences in *legA* as shown in figure 2. In view of the fact that the *legA* sequence shows little homology to the other two gene sequences upstream or downstream of these points it is reasonable to believe that these homologies may have been conserved because of functional constraints.

The perfect sequence homology between all three genes in the region proximal to the start of translation is somewhat surprising. This could indicate that the *legA*, *legB* and *legC* genes have arisen by recent duplications of an ancestral gene. This is further supported by our findings on the coding, intron and 3' flanking sequences of the three genes which show a very high degree of homology, with only a total of 2 silent and 8 amino change nucleotide differences in 1800 bp of coding sequence (J. Gatehouse and R. Croy, unpublished work). However the partial homology in the next adjacent region upstream suggests that there has been evolutionary divergence of the sequences in that region. One possible inference is that pre-existing coding sequences have been replaced by copies of another quite recently in the evolutionary past. Such a proposal is consistent with the molecular drive hypothesis of Dover (24) which predicts that members of a gene family such as the legumin family, will be homogenised by transposition or gene conversion. Examination of figure 2 shows that between positions -321 and -570 the legA sequence differs from the legB and C sequences chiefly as a result of insertions or deletions. The insertions that occur in the legA sequence between -321 and -331 and between -424 and -570 are bounded by short repeats of TCTTT and AGTAA respectively (figure 2). Such short repeats are reminiscent of the footprints of transposable elements and their occurrence is consistent with the possibility of transposition of legumin gene sequences.

The distal halves of all three sequences exhibit a number of direct repeats. Some of the more significant of these are indicated by arrows in figure 2. It is apparent from the dot matrix 'self-comparison' of the legC sequence as shown in figure 3C, that it is particularly rich in such sequences. A sequence of 31bp is repeated imperfectly three times between positions -1305 and -1488. This 31bp sequence is partly homologous to two 82bp tandem repeats that occur between -1105 and -1273, just beyond the point where the legB and legC sequences diverge. These 82bp repeats have an interesting structure which is shown in figure 4. Each of the direct repeats contains within it a pair of inverted repeats which may potentially form stem-loop structures. Since the unpaired region of one loop is complementary to that of the other loop, the whole region is also capable of forming one larger stem-loop structure. The length of the base paired region in this stem-loop suggests that it could be a fairly stable structure. It is possible that this represents one end of, or the remnants of, a transposon, since the ends of transposable elements often exhibit inverted repeats (25). In the legC sequence just at the position where the homology with legB is lost (-1069, Fig. 2) is a short sequence of 17bp (TTTTTGGCGTAGTGCTA) which has strong homology with the 3' ends of a number of transposable elements including the insertion element in the Lel soybean lectin gene (26), the Tam I element of *Antirrhinum* (27) and the En-I element of maize (25). This homology together with the potential secondary structures and the abrupt loss of homology at this

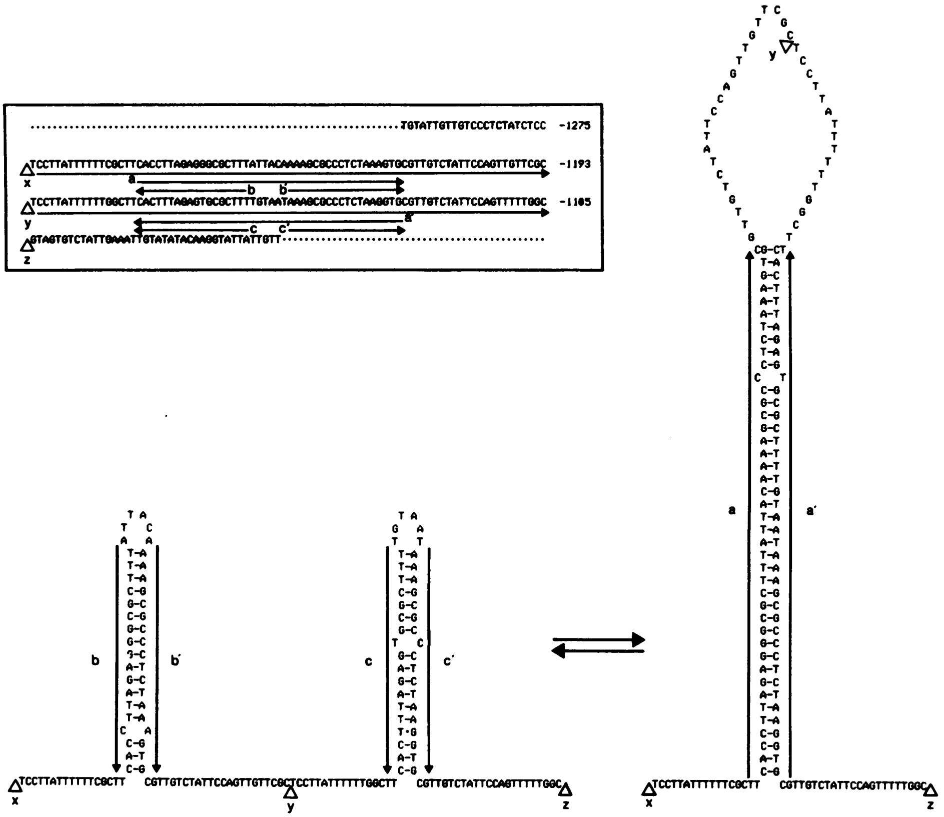

Figure 4 Structural details of the repeated sequences in the legC 5' flanking regions. The extent and direction of the sequence repeats are indicated by the solid arrows labelled a, b, c and d. The potential stem-loop structures which may be formed by the repeat sequences are shown below.

point lends further support to the involvement of pea transposable elements in the generation of sequence diversity in the storage protein genes. This again would be consistent with the proposal that the leg gene family has been subject to a transposon-mediated molecular drive process. Alternatively, in the event that the legC gene is expressed, the structure may be explained as an element involved in gene control. In as much as there is the potential for either of two mutually exclusive structures to be formed, it is analogous to the attenuators that have been found in the 5'-untranslated region of bacterial messenger RNA molecules (28) and more recently in transcripts of the mammalian virus SV40 (29). Although the potential structure in the legC gene is not in the expected place for an attenuator it may perform a different funtctional role by a similar mechanism of switching between two alternative structures. A sequence $(AT)_{22}$ has been located 800bp upstream from this structure (data not shown). Since tracts of alternating purine and pyrimidine bases may form Z-DNA under appropriate conditions (30), any alteration in superhelical density of the DNA as a result of the formation of alternative stem-loops might be offset by switching of the poly-(AT) tract between B- and Z-DNA structures. However, since no such features have so far been found in the upstream sequences of the legA and legB genes it is difficult to envisage what sort of control function, if any, they may play. It is conceivable that the structure lies upstream from a block of legumin genes including legA, legB and legC and modulates coordinate expression, although as yet we have no evidence to support the proposal that the legumin gene family is closely clustered or that all of the members are expressed. Complex stem-loop structures have also been implicated in the control of DNA replication in prokaryotes (31). However, whilst a number of possible functions may be proposed for these unusual inverted repeats in the legC gene, there is no evidence to support any alternative at present.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the excellent technical assistance of Mr. T. Pickard, Mr. J. Cottrell and Mr. D. Jobes and also Dr. J. Gatehouse for useful discussions and suggestions concerning this paper. We are grateful to Dr. D. Baulcombe and Mr. R. Martienssen at the Plant Breeding Institute, Cambridge, for pointing out the legC-transposon homology. We also thank Mrs. E. Ellis for typing the manuscript. This work was supported by the SERC, UK.

[1]Present address: Department of Physiology and Environmental Science, University of Nottingham School of Agriculture, Sutton Bonington LE12 5RD, UK

*To whom correspondence should be addressed

REFERENCES

1. Gatehouse, J.A., Evans, I.M., Bown, D., Croy, R.R.D. and Boulter, D. (1982) Biochem. J. 208, 119-127.
2. Chandler, P.M., Higgins, T.J.V., Randall, P.J. and Spencer, D. (1983) Plant Physiol. 71, 47-54.
3. Chandler, P.M., Spencer, D., Randall, P.J. and Higgins, T.J.V. (1984) Plant Physiol. 75, 651-657.
4. Evans, I.M., Gatehouse, J.A., Croy, R.R.D. and Boulter, D. (1984) Planta 160, 559-568.
5. Lycett, G.W., Croy, R.R.D., Shirsat, A.H. and Boulter, D. (1984) Nucleic Acids Res. 12, 4493-4506.

6. Lycett, G.W., Delauney, A.J., Zhao, W., Gatehouse, J.A., Croy, R.R.D. and Boulter, D. (1984) Plant Mol. Biol. 3, 91-96.
7. Evans, I.M., Bown, D., Lycett, G.W., Croy, R.R.D., Boulter, D. and Gatehouse, J.A. (1985) Planta (In press).
8. Shaw, C.H., Carter, G.H., Watson, M.D. and Shaw, C.H. (1984) Nucleic Acids Res. 12, 7831-7846.
9. Coruzzi, G., Broglie, R., Edwards, C. and Chua, N.-H. (1984) EMBO J. 3, 1671-1697.
10. Gruss, P. (1984) DNA 3, 1-5.
11. Graham, E.E. (1978) Analyt. Biochem. 85, 609-613.
12. Maniatis, T., Fritsch, E.F. and Sambrook, J. (1982) Molecular Cloning: A Laboratory Manual, Cold Spring Harbor Laboratory, New York.
13. Loenen, W.A.M. and Brammar, W.J. (1980) Gene 20, 249-259.
14. Leder, P., Tiemier, D. and Enquist, L. (1977) Science (Wash.) 196, 175-177.
15. Croy, R.R.D., Lycett, G.W., Gatehouse, J.A. and Boulter, D. (1984) Kulturpflanze 32, 81-97.
16. Vieira, J. and Messing, J. (1982) Gene 19, 259-268.
17. Messing, J. and Vieira, J. (1982) Gene 19, 269-276.
18. Seif, I., Khoury, G. and Dhar, R. (1980) Nucleic Acids Res. 8, 2225-2240.
19. Sanger, F., Nicklen, S. and Coulson, A.R. (1977) Proc. Natl. Acad. Sci. USA 74, 5463-5467.
20. Weiher, H., Konig, M. and Gruss, P. (1983) Science (Wash.) 219, 626-631.
21. Gatehouse, J.A. and Croy, R.R.D. (1984) in CRC Critical Reviews in Plant Sciences 1, 287-314, CRC Press Inc.
22. Croy, R.R.D., Lycett, G.W., Gatehouse, J.A., Yarwood, J.N. and Boulter, D. (1982) Nature 295, 76-79.
23. Domoney, C. and Casey, R. (1985) Nucleic Acids Res. 13, 687-699.
24. Dover, G. (1982) Nature 299, 111-117.
25. Gierl, A., Schwartz-Sommer, Z. and Saedler, H. (1985) EMBO J. 4, 579-583.
26. Vodkin, L.O., Rhodes, P.R. and Goldberg, R.B. (1983) Cell 34, 1023-1031.
27. Bonas, U., Sommer, H. and Saedler, H. (1984) EMBO J. 3, 1015-1019.
28. Keller, E.B. and Calvo, J.M. (1979) Proc. Natl. Acad. Sci. USA, 76, 6186-6190.
29. Hay, N., Skolnik-David, H. and Aloni, Y. (1982) Cell 29, 183-193.
30. Arnott, S., Chandrasekaran, R., Birdsall, D.L., Leslie, A.G.W. and Ratliff, R.L. (1980) Nature 283, 743-745.
31. Tomizawa, J. (1983) in Nucleic Acid Research: Future Development, Mizobuchi, K., Watanabe, I. and Watson, J.D. Eds., pp. 475-485, Academic Press, Tokyo, New York and London.
32. Staden, R. (1982) Nucleic Acids Res. 10, 2951-2961.