Nucleotide sequences from phaseolin cDNA clones: the major storage proteins from *Phaseolus vulgaris* are encoded by two unique gene families

Jerry L.Slightom^{*+}, Roger F.Drong⁺, Rodney C.Klassy[§] and Leslie M.Hoffman

Agrigenetics Advance Research Division, 5649 East Buckeye Road, Madison, WI 53716, USA

Received 8 July 1985; Revised 14 August 1985; Accepted 23 August 1985

ABSTRACT

The nucleotide sequences of eight partial and five full-length phaseolin cDNA clones show that phaseolin polypeptides are encoded by two distinct gene families which differ in their coding regions by the presence or absence of two different size direct repeats. The α -type phaseolin polypeptides are encoded by genes containing direct repeats which encode 14 additional amino acids. Aside from these differences, the α - and β -type phaseolin genes show a high degree of homology (98%) which is consistent with these genes being derived from a common ancestral gene. Much of the heterogeneity found in the phaseolin polypeptides appears to be due to post-translational processing. Nucleotide sequence analysis demonstrates that the α -type genes contain only a few amino acid replacement substitutions and that the β -type genes appear to contain no amino acid replacement substitutions. S1 nuclease mapping shows a complex pattern for transcriptional initiation of phaseolin mRNA. Hydropathy analysis shows that phaseolin polypeptides are predominately hydrophilic, and that the two N-glycosyl recognition sites are located in different hydropathic environments.

INTRODUCTION

The seed storage proteins from nutritionally important plants, belonging to the family Leguminosae, have been the subject of many investigations designed to determine their physical characteristics, developmental regulation, post-translational processing, transport, and storage (for reviews see ref. 1 and 2). Generally, the storage proteins found in these seeds are globulin-type (salt-extractable) and consist of two protein types which possess different physical and chemical properties, the glycosylated 7S (vicilin) and non-glycosylated 11S (legumin) types (1). The seeds of French bean (*Phaseolus vulgaris* L.) contain the 7S vicilin-type storage proteins, which are also referred to as phaseolins (3), and constitute about 50% of the total protein in mature seeds (4).

One-dimensional NaDodSO₄/polyacrylamide gel electrophoresis (PAGE) of phaseolins from the cultivar 'Tendergreen' shows three polypeptide bands, α , β , and γ - which have apparent molecular weights of 51-53, 47-48, and 43-46 kilodaltons (kd), respectively (5,6). These polypeptides have similar amino acid compositions and proteolytic and chemical cleavage profiles (7,8) which suggest that they are encoded by a relatively small and conserved multigene family. Gene copy number analyses indicate

that there are approximately seven phaseolin genes per haploid genome (9). Genetic analysis of *P. vulgaris* cultivars which exhibit different two-dimensional PAGE protein patterns show no recombinant protein phenotypes suggesting that these genes are closely linked (10). However, at the molecular level the distance between these genes could be considerable (3,9).

In the developing bean cotyledon phaseolin polypeptides are found to be encoded in a 16S mRNA species which represents about 40% of the total poly(A) RNA (11,12). After translation, the phaseolin polypeptides are involved in a series of processing steps including signal peptide cleavage (13,14) and the stepwise addition (15) of oligosaccharide chains to the two N-glycosyl sites present (13). These post-translational steps appear to be responsible for much of the heterogeneity found among phaseolin polypeptides by two-dimensional PAGE (6,14,15). Evidence supporting this conclusion comes from finding that the signal peptide varies in length between 21 to 25 amino acid residues (14) and that non-complex polypeptide profiles are observed when *in vitro* translated phaseolin polypeptides (which are not glycosylated during translation by a wheat germ system) are analyzed by one-dimensional PAGE (showing only two polypeptide bands 47 and 43 kd [16]) and two-dimensional PAGE (showing only three polypeptides, one β - and two α -types [17]). These results again suggest that the phaseolin multigene family does not vary greatly with respect to charged amino acid replacement substitutions or insertions or deletions which would alter their polypeptide sizes.

In order to more accurately determine the extent of genetic variability among the individual members of the phaseolin multigene family we obtained the nucleotide sequences for eight partial and five full-length phaseolin cDNA clones. Partial nucleotide sequences for several of these full-length clones have been reported previously (13,18) and in the present report we have analyzed the complete nucleotide sequences from eight partial and all five full-length cDNA clones. This analysis shows that phaseolin polypeptides are encoded in two distinct phaseolin gene types or subfamilies. S1 nuclease protection was used to determine the authenticity of the 5' termini capped nucleotides found in the presumed full-length phaseolin cDNA clones. Hydropathic profiles for both α - and β -type phaseolin polypeptides were also obtained.

MATERIALS AND METHODS

Materials

Restriction endonucleases: Bgl I, Eco RI, Nco I, Pst I, Sac I, and Xba I were obtained from Promega-Biotech or New England Biolabs. Polynucleotide kinase was from P-L Biochemicals and bovine alkaline phosphatase was from Boehringer-Mannheim. [γ -³²P] ATP (2000-3000 Ci/mM) and [α -³²P] dCTP (800 Ci/mM) were obtained from New

England Nuclear. Chemicals used for DNA sequencing were obtained from the vendors recommended by Maxam and Gilbert (19). X-ray film on rolls (20 cm x 25 m), XAR-351, was obtained from Kodak. Intensifying screens, Quanta III (35 cm x 1 m) from DuPont were cut in half to fit sequencing gels. DNA sequencing gel stands, designed for gels measuring 20 cm x 104 cm, and safety cabinets were designed by J. Slightom, R. Barker, and J. Smoot and are available from Fotodyne, Inc., New Berlin, Wisconsin. Water jacket thermostating plates were constructed in our laboratory using 1/4 inch thick plate glass glued together by 100% silicone cement.

Methods

mRNA Purification and cDNA Cloning. Total RNA was prepared from French bean cotyledons ('Tendergreen') according to the method described by Hall *et al.* (16), and poly(A) RNA was obtained by poly(U) Sephadex chromatography (20). Double-stranded cDNA was prepared by the method of Land *et al.* (21), or by using a modification to this method described by Murray *et al.* (22). C-tailed cDNAs were cloned by annealing with G-tailed (at the *Pst* I site) pBR322 followed by transformation into *E. coli* HB101 as described by Dagert and Ehrlich (23). Recombinant phaseolin cDNA clones were screened using the procedure described by Grunstein and Hogness (24) using the insert from a previously isolated phaseolin cDNA clone, cDNA1(3), as ³²P-labeled probe. DNA was isolated from these clones using the procedure of Birnboim and Doly (25), followed by two CsCl-ethidium bromide gradient bandings.

DNA Sequence Determination. DNA sequences were determined using the chemical method of Maxam and Gilbert (19). Sequenced DNAs were electrophoresed 14 hours at 2,500 constant volts on gels measuring 20 cm wide, 104 cm long, and 0.2 mm thick as described by Barker *et al.* (26). Constant temperature (50°C) was maintained using a water-jacketed plate on one side of the gel sandwich. The opposite plate of the sandwich was treated with γ -methacryloxypropyltrimethylsilane (Sigma 6514) as described by Garoff and Ansorge (27) so that the acrylamide used to form the gel matrix is bonded directly to the glass. Gel pouring, loading, and autoradiography conditions have been described previously by Slightom *et al.* (13).

S1 Nuclease Mapping. The location of phaseolin transcription initiation sites was determined using the S1 nuclease mapping procedure (28). A 550 bp *Neo* I to *Eco* RI fragment, containing 5' flanking, untranslated and coding regions of the phaseolin gene in clone AG-pPvPh 3.8 (13) was 5'-labeled at the *Eco* RI site, denatured, and hybridized with 5 μ g of bean cotyledon total poly(A) RNA for 6 hours at 52°C as described by Hoffman (29). Unprotected mRNA and single stranded DNA were removed by adding 100 μ l of S1 nuclease buffer, pH 5.5 (30) and between 50 to 100 units of S1 nuclease (Sigma, one unit of S1 nuclease renders 1 μ g of single-stranded DNA acid soluble in 1 minute at 37°C). S1 nuclease digestions were also done using 200 units of S1 nuclease

and identical results were obtained. Digests were incubated for 30 minutes at 37°C, extracted with phenol, ethanol precipitated, and run on a 6% sequencing gel along with the sequenced Nco I to Eco RI fragment.

Computer Analyses of DNA Sequences. The following computer programs, supplied by the University of Wisconsin Genetics Computer Group (31), were used: COMPARE, DOT PLOT, MAP and STEMLOOP. The hydrophobicity index plot is based on a program by Kyte and Doolittle (32) and modified for use on a Pyramid 90X computer by C. Brown (Agrigenetics Advanced Research Division). Polypeptide isoelectric points were calculated using computer programs provided by DNASTAR, Madison, Wisconsin.

RESULTS AND DISCUSSION

Isolation of Phaseolin cDNA Clones

Several attempts have been made to obtain full-length phaseolin cDNA clones from purified 16S phaseolin mRNA. Phaseolin mRNA-enriched fractions were obtained from mid-maturation stage 'Tendergreen' cotyledon poly(A) RNA after sucrose gradient centrifugation. *In vitro* translation of mRNA gradient fractions indicate that the 16S mRNA fraction encodes phaseolin (3,16). Our first attempt to construct phaseolin cDNA clones from this 16S mRNA fraction resulted in the isolation of clone cDNA1, which contains a phaseolin cDNA insert of about 570 bp arising from the center of the phaseolin coding region (3; see Fig. 1). Two additional attempts to isolate full-length phaseolin cDNA clones have been described by Murray *et al.* (22) and utilized the cDNA method described by Land *et al.* (21) and a modification of that procedure. Partial phaseolin cDNA clones 55, 113, and 198 were prepared using the Land *et al.* (21) procedure, and the longer cDNA clones 6, 13, 31, 39, 69, 72, 110, 134, and 169 were obtained using the modified procedure. All of these cDNA clones show the same pattern of restriction sites, indicating that they are internally homologous (see Fig. 1).

Nucleotide Sequence of Phaseolin cDNA Clones

Because phaseolin cDNA inserts contain identical restriction enzyme sites, the same sequencing strategy (Fig. 2) was applicable for all clones. The nucleotide and derived amino acid sequences for these nine longest clones are presented in Fig. 3. Clearly, all of the sequences share a high degree of homology and several of these cDNA sequences are identical (cDNA clones 39 and 134, and cDNA clones 69 and 169). Also, cDNA clones 31 and 72 appear to be identical except for a one base pair deletion in clone 72 (position 179). This deletion results in a frame shift which encounters a translation termination signal (TAG) 105 bp downstream and thus this transcript cannot encode a phaseolin polypeptide. Such a deletion could have occurred during the cDNA cloning experiment, or this clone could represent a phaseolin gene which has acquired a deletion, thus becoming a pseudogene. The latter case could only be substantiated with

Fig. 1. Size, restriction enzyme site maps, and classifications of phaseolin cDNA clones. The maps of thirteen phaseolin cDNA inserts cloned into the *Pst* I site of pBR322 are presented. The restriction enzyme sites shown above the size indicator line were found in similar positions in all the cDNA clones and were used to determine the region of the phaseolin transcript contained in each clone. The exact nucleotide length of each cDNA clone was determined by nucleotide sequencing and is shown to the right. These cDNA clones have been grouped according to the type of phaseolin polypeptide encoded, either α - or β -type (see Fig. 3). The triangles below the α -type clones show the positions of the direct repeats which distinguish the α - and β -type clones (see text).

the isolation of a phaseolin gene containing this exact deletion. Clone 72 also contains an abnormal 3'-untranslated region containing an additional 39 bp (see Fig. 3). These 39 bp appear to be derived from the adjacent 3'-flanking region of a phaseolin gene because its sequence shares 90% homology with the 3'-flanking DNA of the phaseolin gene cloned in AG- λ PvPh177.4 (J.L. Slightom, unpublished results).

Lack of complete sequence identity among the remaining phaseolin cDNA sequences is mostly due to the presence of a hypervariable region located between

Fig. 2. Nucleotide sequencing strategy used for full-length phaseolin cDNA clones. Because all of the cDNA clones shown in Fig. 1 contained the same restriction enzyme sites, the same sequencing strategy was used. The horizontal arrows show which enzyme sites were used for sequencing and the directions and distance in which sequences were obtained.

Table 1 Summary of Full-Length Phaseolin cDNA Sequences

	α -6	α -39	α -134	β -31	β -72 ⁺
Insert sizes	1535 bp	1517 bp	1514 bp	1475 bp	1511 bp
5'-untranslated Regions	89 bp	80 bp	73 bp	80 bp	73 bp
3'-untranslated Regions	141 bp	136 bp	136 bp	135 bp	173 bp
Coding Regions	1305 bp	1302 bp	1302 bp	1260 bp	1260 bp
Signal Peptide* (amino acids)	23	23	23	23	23
Mature Polypeptide (amino acids)	412	411	411	397	397
Calculated Molecular Weight (Daltons)	46,700	46,600	46,600	45,000	45,000

* Size of signal peptide assumes translation initiation at the first Met residue and cleavage after the Ala residue yielding Thr as the N-terminus of the mature polypeptide.

+ The nucleotide deletion at position 179 is not recognized for the purpose of this table (see text).

positions 456 to 462 (Fig. 3). The nine cDNA clone sequences (Fig. 1) appear to represent seven unique phaseolin genes. Whether these seven unique phaseolin transcripts are the result of nucleotide substitutions in allelic genes or divergence of non-allelic genes is presently unknown. [However, because beans are inbred the amount of heterozygosity should be small (although this has not yet been determined), and these transcripts could represent the expression of seven non-allelic phaseolin genes.] This number of non-allelic phaseolin transcripts is consistent with the gene copy number derived from solution hybridization kinetics and genomic nitrocellulose blot hybridization (9).

Phaseolin Polypeptides are Encoded by Two Distinct Gene Families

A striking feature revealed by comparing the sequences shown in Fig. 3 is the presence or absence of two different size direct repeats in the coding DNA. Direct repeats appear to be a common feature among storage proteins and are responsible for much of the size heterogeneity seen in pea legumin (33) and in maize zein (34). The first direct repeat (15 bp) occurs between positions 711 and 740, and the second direct repeat (27 bp) is located between positions 1322 to 1375. From similar analyses of length differences between globin gene sequences (35,36), it has been suggested that DNA length differences are probably the result of duplications if their lengths are

is noted in any of these sequences, nucleotides from all the cDNA clones are shown for that position. The amino acid sequence derived from cDNA clone 6 is shown below the nucleotide sequence lines, and, if nucleotide substitutions or insertion/deletions alter this amino acid sequence, such changes are shown below the appropriate amino acid residues. Intron positions found in clone λ AG-PhPv177.4 (13) are indicated by vertical arrows; however, it is not known if these positions are conserved throughout the other members of the phaseolin gene families. Phaseolin cDNA clones which contain the 5'-capped terminal nucleotide are indicated and have been confirmed (see Fig. 4). Nucleotides which were polyadenylated were determined from nucleotide sequences. Horizontal arrows above the nucleotide sequence of clone 6 show positions of the direct repeats which distinguish the α - and β -type gene families. Horizontal arrows shown below the sequences indicate the location of 5'-untranslated DNA, signal peptide, or 3'-untranslated DNA. The vertical arrows in the amino acid sequence near the signal peptide show the multiple cleavage positions found for removal of the signal peptide (14). Double underlined amino acids residues denote possible N-glycosyl attachment recognition sites.

equivalent to the length of the nucleotide gap. However, if DNA length differences are separated by an unspecified number of nucleotides they are probably the result of deletions. These criteria suggest that the α -phaseolin gene has evolved via two duplication events (which could have occurred separately or simultaneously) and several point mutations from an ancestral β -type phaseolin gene.

Derived polypeptide molecular weights were calculated for the different classes of phaseolin transcripts, see Table 1. These calculations have been made assuming cleavage of an identical signal peptide after the Ala residue (Fig. 3), yielding Thr as the N-terminal amino acid (14). The 1.8 kd difference (14 amino acids) between the larger α -type and the smaller β -type phaseolin polypeptides is explained by the presence or absence of the direct repeats mentioned above. Thus, cDNA clones 6, 13, 39, 110, 134, and 169 encode α -type phaseolin polypeptides and cDNA clone 31 and 72 encode β -type polypeptides. The shorter cDNA clones (1, 55, 113, and 198) are probably derived from β -type phaseolin transcripts because they do not contain either direct repeat. We presently have no evidence indicating that a phaseolin gene exists containing either the 15 or 27 bp direct repeat, which suggests that these direct repeats may have evolved simultaneously in *P. vulgaris* and do not segregate (see below). These results suggest that phaseolins are encoded by two unique but highly homologous gene subfamilies which appear to resemble closely the two gene families encoding the related soybean storage proteins α and α' subunits of β -conglycinin (37).

Glycosylation Plays a Role in Phaseolin Size Heterogeneity

We found no cDNA clone transcript sequence that could account for the γ -type phaseolin polypeptide which is expected to be about 1.8 kd smaller than the β -type polypeptide (5,6). Also, nucleotide sequences from three phaseolin genes arising from the three unique Eco RI fragment size classes described by Talbot *et al.* (9) (6.7, 7.2, and 13.0 kb fragments classes) show coding sequences consistent with only the α - and β -type

sequences shown in Fig. 3 (13; J. Slightom, D. Thompson, and R. Drong, in preparation). Therefore, the existence of a gene which encodes the γ -type polypeptide is doubtful. The γ -type polypeptide appears to be the result of a post-translational event, namely partial glycosylation of a β -type polypeptide backbone. This conclusion is supported by *in vitro* translation of the 16S mRNA which yields only two phaseolin polypeptides, of 43 and 47 kd (15,16,17) and because the γ -polypeptide shows a smaller carbohydrate to protein ratio than either the α - or β -type polypeptides (14,15). The γ -polypeptide may be the result of a single glycosylation event or oligosaccharide processing of the β -type polypeptide (15). Both α - and β -type polypeptides contain two identical N-glycosyl recognition sites, thus differential glycosylation or oligosaccharide processing could be responsible for much of the size heterogeneity found in both types of phaseolin polypeptides (6,14,15,38). The degree that oligosaccharide variations contribute to both α - and β -type polypeptide heterogeneity appears to be considerable (17), much more than can be accounted for by amino acid replacement substitutions (see below). Bollini *et al.* (15) suggest that the addition of oligosaccharide side chains is a complex process which possibly involves degradation or modification of the oligosaccharide core, and this process may be affected by the different protein environments surrounding the two N-glycosyl recognition sites (see below).

Amino Acid Substitutions in the α - and β -Type Phaseolin Polypeptides

Two-dimensional PAGE analysis of *in vitro* translated (non-glycosylated) phaseolin polypeptides (from cultivar 'Tendergreen') reveals a very simple polypeptide profile consisting of only two α -type polypeptides separated by charge and a single cluster of β -type polypeptides separated from the α -type polypeptides by both charge and size (17). This result supports the claim (see above) that the oligosaccharide sidechains are responsible for much of the heterogeneity found among these polypeptides, and it also indicates that a limited amount of amino acid replacements exist among these polypeptides which alter their isoelectric points. The data presented in Fig. 3 (and also from the partial β -type clones) is consistent with this observation because no amino acid replacement substitutions have yet been found in the β -type gene sequences while the α -type genes show only two potential charge variations; a sequence which contain one additional amino acid, Gln, (in α -6, the 412 residue polypeptide) and one charged amino acid replacement substitution in the 411 residue polypeptides, Asn, (in α -13, -39 and -134 clones) or, His, (in α -69, -110, and -169 clones) see Fig. 3 position 486. The calculated isoelectric points for the non-glycosylated α -type polypeptides are, pH 5.10 for both polypeptides containing the additional Gln and Asn residues while the α -type polypeptide containing the His replacement residue has its isoelectric point at pH 5.16. All β -type polypeptides have a calculated isoelectric point of pH 5.13. The two-dimensional PAGE profile expected from these non-glycosylated polypeptides is exactly

that observed by Lioi and Bollini (17), which suggests that the sequences in Fig. 3 account for all the expected charge variations.

Substitution Pattern Among the Phaseolin Gene Families Indicate a Homogenization Mechanism

The number of nucleotide substitutions among the respective members of the α - and β -type phaseolin gene families (excluding the hypervariable region, positions 456-462) is quite small with no substitutions found among β -type genes and only four substitutions found among the α -type genes, three of which result in amino acid replacements (Fig. 3). The number of nucleotide substitutions between the α - and β -type phaseolin gene families is also small showing only 16 substitutions. These two gene families differ by less than 2%. Eleven of these substitutions are in the coding DNA, nine of which result in amino acid replacements. The remaining five substitutions are located in the 3' untranslated region, three of which are contiguous (see positions 1452-1565). It is interesting to note that nucleotide differences among intra- and intergenic family members appear to favor amino acid replacement rather than silent substitutions. The presence of these amino acid replacement substitutions and the two regions of amino acid length variations indicate that there are parts of the phaseolin polypeptide that can be altered without adversely affecting its biological function. No nucleotide substitutions are located in the 5'-untranslated region (although this region does vary in length, see below), the signal peptide, and the first 59 amino acids. Sequence conservation in these regions suggest that they have biological importance, most likely because they are involved in translation initiation and signal peptide cleavage, respectively.

The substitution pattern between the α - and β -phaseolin gene families is also interesting because up to and including position 596, the substitution patterns are mixed (switching between α - and β -type sequences) and after that position, the substitutions show no mixing. The mixed substitutions are in positions 339, 486, and 596. Only one substitution among these sequences is unique and that occurs in cDNA clone α -6 at position 533. Conservation of this high degree of nucleotide sequence homology among seven possibly non-allelic phaseolin genes suggests that both inter- and intragene family homogenization, similar to that found in many other multigene families, may be occurring. If recombinational mechanisms are responsible for homogenization of these gene families, the present data indicate that the two direct repeat sequences (see Fig. 3) do not segregate as a result of intragene family exchanges in the 'Tendergreen' cultivar. Recombination mechanisms usually involve crossover and/or gene conversion events (39,40) occurring between identical gene sequence regions. Determination of the amount of DNA involved and the frequency of these recombinational events requires additional gene linkage and sequence information. A knowledge of the physical linkage

Fig. 4. Localization of phaseolin transcription initiation sites by S1 nuclease mapping. A 550 bp *Nco* I-*Eco* RI fragment spanning the phaseolin promoter region was 5'-end labeled at the *Eco* RI site and reassociated with tRNA (lane 1) or bean cotyledon poly(A) RNA (lane 2). Samples were digested with 100 units of S1 nuclease and protected DNA fragments were electrophoresed in a 6% polyacrylamide sequencing gel along with the sequenced *Nco* I-*Eco* RI fragment (see nucleotide sequence lanes located to the right). Nucleotides used for transcription initiation are indicated by arrows offset by a 1.5 base correction (43). Capped nucleotides found in the phaseolin cDNA clones are shown by their corresponding clone numbers to the right of the printed nucleotide sequence.

arrangement of the phaseolin genes would be useful in determining the importance that recombinations play on the evolution of these gene families. Presently, we do not know whether the α - and β -type phaseolin genes are arranged as pure or interdispersed gene families. Our sequence data suggest that the β -type genes may be arranged as a closely linked pure family, showing no substitutions, because they have co-evolved (due to frequent recombinations) with neighboring β -type genes. In contrast, the α -type gene family show more variability and may not be as closely linked to each other as the β -type genes and thus may not be involved in frequent recombinations. DNA hybridization studies (9,12) indicate that the different members of each gene families are homologous

Fig. 5. Hydropathic profile of phaseolin protein sequence derived from cDNA clone 6. Hydrophobic domains are indicated by a positive index and hydrophilic domains are indicated by a negative index. The large arrow indicates the major position for signal peptide cleavage. The location of the two direct repeats found in clone cDNA 6 are shown by lines below the index and the position of the two N-glycosyl addition sites are also shown (CHO).

outside of the transcript region, in particular the region extending from the cap site in the 5'-direction (upstream) for about 800 bp. Interestingly, this 800 bp of upstream DNA is contained in the phaseolin T-DNA constructions described by Murai *et al.* (41) and this DNA region has been shown to contain the regulatory elements necessary to control developmental expression of phaseolin in tobacco seeds (42).

S1 Nuclease Mapping of Phaseolin Transcriptional Starts

The 5'-terminal nucleotides of the cDNA clones shown in Fig. 3 are expected to represent the 5'-capped nucleotide of their respective transcript. The position of the capped nucleotide found in these phaseolin cDNA clones shows considerable variability (see Fig. 3). We used the S1 nuclease mapping procedure of Weaver and Weissman (28) to substantiate whether the 5'-terminal nucleotides shown in Fig. 3 actually represent the 5'-capped nucleotides. Cotyledon poly(A) RNA was hybridized with a 550 bp Nco I-Eco RI fragment, isolated from the genomic clone AG-pPvPh3.8 (13), which had previously been 5' end-labeled at the Eco RI site. Because this Eco RI site is located in all the cDNA clones, see Fig. 3, and all of the phaseolin genes are homologous over this stretch of DNA, hybridizations using this Nco I-Eco RI fragment should reflect the contributions from all the phaseolin transcripts. A very complex arrangement of S1 nuclease protected DNA fragments was revealed, see Fig. 4. There appear to be at least twelve S1-protected DNA bands arranged in two clusters separated by 5 to 6 bp.

The lower cluster is predominate and represents eight DNA fragments, while the upper less intense cluster represents four DNA fragments. Allowing for a 1.5 base correction (43), the 5'-termini found in the cDNA clones have been identified (Fig. 4). Seven of the nine cDNA clones presented in Fig. 3 appear to contain the 5' capped nucleotides while the other two cDNA clones, 13 and 110, terminate prematurely. Six of the cDNA clones have transcription initiation starting from the lower cluster while only cDNA clone 6 was initiated from the upper cluster.

Varability in transcription initiation is not unusual (29) but generally involves only two to three nucleotides near a cap site. The presence of possibly twelve transcription initiation nucleotides is quite unexpected. However, as our sequencing results show, at least three of these nucleotides are utilized (Fig. 4). A possible explanation for this result was observed by Slightom *et al.* (13) who found that the 5' flanking DNA of the phaseolin gene AG- λ PvPh177.4 contains multiple and overlapping TATA-elements. The presence and arrangement of these multiple TATA-elements appears to be the major contributor to the appearance of these multiple cap sites rather than the fact that these phaseolin transcripts arise from a multigene family. Evidence supporting this hypothesis come from two sources. First, the two clusters of S1 protected fragments in Fig. 4 are separated by about 5 to 6 bases which is identical to the spacing between the TATA-elements found in the 5' flanking DNA (13). Secondly, S1 nuclease mapping of transcripts derived from the single phaseolin gene transferred into sunflower tissue shows the same complex pattern of S1 protected fragments (41).

Hydropathy analysis of phaseolin polypeptides and its influence on post-translational modification

Hydropathic profiles for the polypeptides encoded by cDNAs 6 and 31 were obtained using a modification of the computer program described by Kyte and Doolittle (32). Fig. 5 shows that cDNA-6 encodes a protein which is predominately hydrophilic in nature with its most intensely hydrophobic region corresponding with the signal peptide (13,14). The direct repeats present in the α -type polypeptides do not alter its hydropathy profile significantly from the β -type profile. It is interesting to note that the two N-glycosyl recognition sites are located in regions with very different hydropathic environments (Fig. 5). Analysis of oligosaccharide addition to phaseolin by Bollini *et al.* (15) shows that two oligosaccharide chains can be added co-translationally to both α - and β -type phaseolin polypeptides and that their addition involves two steps. The different hydropathic environments in which each N-glycosyl recognition site is located may be responsible for the two step glycosylation of phaseolin polypeptides. The glycosylation site in exon 4 is located in a hydrophilic region and may be more accessible for glycosylated during translation than the site in exon 5 which is located in a hydrophobic region. Differential glycosylation could occur at either of these sites, but

because the glycosylation site in exon 5 sets in a hydrophobic domain we suggest that its glycosylation may involve a more complex process (which may depend on protein unfolding) thus its location could be responsible for some of the α - and β -type polypeptide heterogeneity.

ACKNOWLEDGEMENTS

We are grateful to Dr. Mike Murray for constructing full-length phaseolin cDNA clones, to Conrad Brown and Jeff Lotzer for help with computer analysis, and to Jo Adang for illustrations and photographs. This is Agrigenetics Advanced Research Division Manuscript No. 39.

Present addresses: §Texas A & M University, Department of Biology, College Station, TX 77843 and
 +The Upjohn Company, Kalamazoo, MI 49001, USA

*To whom correspondence should be addressed

REFERENCES

1. Derbyshire, E., Wright, D.J. and Boulter, D. (1976) *Phytochem.* 15, 3-24.
2. Ersland, D.R., Brown, J.W.S., Casey, R. and Hall, T.C. (1983) In: Gottschalk, W. and Müller, H.P. (eds) *Seed proteins: Biochemistry, genetics, nutritive value.* Martinus Nijhoff/Dr. W. Junk, The Hague pp. 355-375.
3. Sun, S.M., Slightom, J.L. and Hall, T.C. (1981) *Nature* 289, 37-41.
4. Ma, Y. and Bliss, F.A. (1978) *Crop Science* 17, 431-437.
5. Hall, T.C., McLeester, R.C. and Bliss, F.A. (1977) *Plant Physiol.* 59, 1122-1124.
6. Brown, J.W.S., Ma, Y., Bliss, F.A. and Hall, T.C. (1981) *Theor. Appl. Genet.* 59, 83-88.
7. Ma, Y., Bliss, F.A. and Hall, T.C. (1980) *Plant Physiol.* 66, 897-902.
8. Bollini, R. and Vitale, A. (1981) *Physiol. Plant* 52, 96-100.
9. Talbot, D.R., Adang, M.J., Slightom, J.L. and Hall, T.C. (1984) *Mol. Gen. Genet.* 198, 42-49.
10. Brown, J.W.S., Bliss, F.A. and Hall, T.C. (1981) *Theor. Appl. Genet.* 60, 251-258.
11. Hall, T.C., Sun, S.M., Buchbinder, B.U., Pyne, J.W., Bliss, F.A. and Kemp, J.D. (1980) In: *Genome Organization and Expression in Plants*, Leaver, C.J., Ed., Plenum Publishing Corporation pp. 259-272.
12. Murray, M.G. and Kennard, W.C. (1984) *Biochem.* 24, 4225-4232.
13. Slightom, J.L., Sun, S.M. and Hall, T.C. (1983) *Proc. Natl. Acad. Sci. USA* 80, 1897-1901.
14. Paaren, H.E., Slightom, J.L., Hall, T.C., Inglis, A.S. and Blagrove, R.J. (1985) *Phytochemistry* Submitted.
15. Bollini, R., Vitale, A. and Chrispeels, M.J. (1983) *J. Cell. Biol.* 96, 999-1007.
16. Hall, T.C., Ma, Y., Buchbinder, B.U., Pyne, J.W., Sun, S.M. and Bliss, F.A. (1978) *Proc. Natl. Acad. Sci. USA* 75, 3196-3200.
17. Lioi, L. and Bollini, R. (1984) *Plant Molec. Biol.* 3, 345-353.
18. Hall, T.C., Slightom, J.L., Ersland, E.R., Murray, M.G., Hoffman, L.M., Adang, M.J., Brown, J.W.S., Ma, Y., Matthews, J.A., Cramer, J.H., Barker, R.F., Sutton, D.W. and Kemp, J.D. (1983) In: *Structure and Function of Plant Genomes*, Ciferri, O. and Dure, III, L., Eds., Plenum Publishing Corporation pp. 123-142.
19. Maxam, A.M. and Gilbert, W. (1980) *Methods in Enzymol.* 65, 499-560.
20. Murray, M.G., Peters, D.L. and Thompson, W.F. (1981) *J. Mol. Evol.* 17, 31-42.
21. Land, H., Grez, M., Hanson, H., Lindenmaier, W. and Schutz, G. (1981) *Nucl. Acids Res.* 9, 2251-2266.

22. Murray, M.G., Hoffman, L.M. and Jarvis, N.P. (1983) *Plant Molec. Biol.* 2, 75-83.
23. Dagert, M. and Ehrlich, S.D. (1979) *Gene* 6, 23-28.
24. Grunstein, M. and Hogness, D.S. (1975) *Proc. Natl. Acad. Sci. USA* 72, 3961-3965.
25. Birnboim, H.C. and Doly, J. (1979) *Nucl. Acids Res.* 7, 1513-1523.
26. Barker, R.F., Jarvis, N.P., Thompson, D.V., Loesch-Fries, L.S. and Hall, T.C. (1983) *Nucl. Acids Res.* 11, 2881-2891.
27. Garoff, H. and Ansorge, W. (1983) *Anal. Biochem.* 115, 450-457.
28. Weaver, R.F. and Weissmann, C. (1979) *Nucl. Acids. Res.* 7, 1175-1193.
29. Hoffman, L.M. (1984) *J. Molec. Appl. Genet.* 2, 447-453.
30. Maquat, L.E., Kinniburgh, A.J., Rachmilewitz, E.A. and Ross, J. (1981) *Cell* 27, 543-553.
31. Devereaux, J., Haeblerli, P. and Smithies, O. (1984) *Nucl. Acids Res.* 12, 387-395.
32. Kyte, J. and Doolittle, R.F. (1982) *J. Mol. Biol.* 157, 105-132.
33. Lycett, G.W., Delauney, A.J., Zhao, W., Gatehouse, J.A., Croy, R.R.D. and Boutler, D. (1984) *Plant Molec. Biol.* 3, 91-96.
34. Pedersen, K., Devereux, J., Wilson, D.R., Sheldon, E. and Larkins, B.A. (1982) *Cell* 29, 1015-1026.
35. Efstratiadis, A., Posakony, J.W., Maniatis, T., Lawn, R.M., O'Connell, 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) *Cell* 21, 653-668.
36. Shen, S., Slightom, J.L. and Smithies, O., (1981) *Cell* 26, 191-203.
37. Schuler, M.A., Schmitt, E.S. and Beachy, R.N. (1982) *Nucl. Acids Res.* 10, 8225-8244.
38. Hubbard, S.C. and Ivatt, R.J. (1981) *Ann. Rev. Biochem.* 50, 555-583.
39. Zimmer, E.A., Martin, S.L., Beverly, S.M., Kan, Y.W. and Wilson, A.C. (1980) *Proc. Natl. Acad. Sci. USA* 77, 2158-2162.
40. Slightom, J.L., Blechl, A.E. and Smithies, O. (1980) *Cell* 21, 627-638.
41. Murai, N., Sutton, D.W., Murray, M.G., Slightom, J.L., Merlo, D.J., Reichert, N.A., Sengupta-Gopalan, C., Stock, C.A., Barker, R.F., Kemp, J.D. and Hall, T.C. (1983) *Science* 222, 476-482.
42. Sengupta-Gopalan, C., Reichert, N.A., Barker, R.F., Hall, T.C. and Kemp, J.D. (1985) *Proc. Natl. Acad. Sci. USA* 82, 3320-3324.
43. Sollner-Webb, B. and Reeder, R.H. (1979) *Cell* 18, 485-499.