Gene, 52 (1987) 41–49
Elsevier

GEN 01915

Multiple variability in the sequence of a family of maize endosperm proteins

(Recombinant DNA; nucleotide sequencing; hybrid selection; Northern analysis; glutelin; zein)

Salomé Prat, Lluis Pérez-Grau and Pere Puigdomènech

Departamento de Genética Molecular, Centro de Investigación y Desarrollo, CSIC, Barcelona (Spain)

Received 21 July 1986
Revised 9 September 1986
Accepted 27 November 1986

SUMMARY

A collection of cDNA clones, corresponding to a group of maize endosperm proteins classified in the glutelin-2 (or reduced soluble proteins) and in the zein-2 subfractions, has been identified and characterized. The nucleotide sequence of three of these clones has been obtained and the amino acid sequence deduced. They appear to correspond to a small family of genes that are specifically expressed in immature endosperm simultaneously to zeins, the best characterized proteins from this tissue. Unlike zeins, the proteins of the glutelin-2 and zein-2 family contain sequences homologous to storage proteins from other cereals such as gliadins or hordeins. The cDNA clones encoding for the two types of proteins have been compared, and a high degree of homology has been observed for both the nucleotide and amino acid sequences. The differences existing in both the coding and non-coding regions allow the definition of multiple types of variability in their sequence. An hypothesis is proposed on how sequence diversity may have been generated in this particular class of plant proteins.

INTRODUCTION

One of the most extensively studied groups of plant genes are those expressed in cereal seeds. The reasons for this may be the interest in the cereal grain proteins for the nutrition of man and animals, the relative ease of specific mRNA preparation and the fact that these genes constitute, particularly in maize, an excellent model to study the mechanisms of gene variability and expression in plants.

The most abundant proteins in seeds are considered to have a storage function. A large group of cereal storage proteins form a superfamily of proteins as deduced from the high degree of homology observed in their sequence. This superfamily of proteins may also include other seed proteins such as the barley trypsin inhibitor and some storage proteins from dicotyledons (Kreis et al., 1985). In contrast, zeins, the main maize storage proteins, show

Correspondence to: Dr. P. Puigdomènech, Departamento de Genética Molecular, Centro de Investigación y Desarrollo, CSIC, Jorge Girona Salgado, 18, 08034 Barcelona (Spain) Tel. (34-3)2046851.

Abbreviations: aa, amino acid(s); cDNA, DNA complementary to mRNA; DEAE, diethylaminoethyl; kb, 1000 base pairs; nt, nucleotide(s); RSP, reduced soluble proteins; SSC, 0.15 M NaCl, 0.015 M Na · citrate, pH 7.

no homology in their sequence with other seed proteins. Nevertheless, the second most abundant protein fraction of maize endosperm, glutelin, is related to storage proteins from other cereals (Prat et al., 1985).

When proteins from maize flour are extracted and analyzed by polyacrylamide gel electrophoresis (Soave and Salamini, 1984) the main groups of bands observed run at 28 kDa (glutelin-2, γ-zeins or RSP), at 22–19 kDa (zeins) and at 14–16 kDa and 10 kDa (zein-2 proteins). Zeins account for approx. 60% of total endosperm protein. Glutelins are the next most abundant group in maize seeds, accounting for around 15% of total protein. These proteins are extracted in alkaline solutions containing a reducing agent, they accumulate at a precise stage of endosperm maturation (between 15 and 30 days after pollination) and they are located, as zeins, in the protein bodies, the storage vesicles of maize endosperm (Ludevid et al., 1984).

The results presented here, obtained by cDNA cloning, indicate that proteins previously classified as glutelins or as components of the 14–16-kDa zein-2 subfraction by the classical sequential extraction method, are encoded by a small family of genes. They have homology with storage proteins from other cereals but they are different from zeins. The comparison of sequences of proteins belonging to this family allows us to define different types of sequence divergence that may correspond to distinct mechanisms of generation of sequence diversity in this type of cereal genes.

MATERIALS AND METHODS

(a) Isolation of RNA, cDNA synthesis and transformation

Total polysomal RNA was extracted from 20-day-old developing *Zea mays* kernels according to Wienand and Feix (1978). Double-stranded cDNA was prepared essentially following the Okayama and Berg (1982) procedure according to Gubler and Hoffman (1983). Following second-strand synthesis, double-stranded cDNA was fractionated on 1.5% agarose gel, and the DNA migrating between 0.5 and 2.0 kb was selected by electroelution onto DEAE-cellulose paper (Dretzen et al., 1981). The double-stranded cDNA was inserted at the *Pst*I site of pBR322 by dG/dC tailing (Villa-Komaroff et al., 1978). *Escherichia coli* HB101 cells were made competent and transformed as described by Hanahan (1983).

(b) Colony hybridization and hybrid selection

Recombinant clones were screened by colony hybridization using as probe the nick-translated (Rigby et al., 1977) ^{32}P-labeled inserts from partial cDNA clones pME117 and pME119 (Prat et al., 1985). Following hybridization, filters were extensively washed in 0.5 × SSC at 65°C, air-dried and autoradiographed.

Positive clones containing the longest cDNA insert were chosen for hybrid selection of specific mRNA. This experiment as well as immunoprecipitation and polyacrylamide gel electrophoresis of in vitro translated products and fluorography were performed as previously described (Prat et al., 1985). The main difference with previous experiments was that filters containing the different plasmids were independently incubated with the RNA to avoid the possibility of sequestering of specific mRNA by any of the inserts.

(c) Northern blot analysis

RNA samples from the W64 and W64O2 inbred lines were separated in 1.5% agarose-formaldehyde denaturing gels and transferred to nitrocellulose paper for hybridization. Hybridizations were performed with cDNA probes labeled by nick-translation. Quantification of autoradiograms was performed by measuring the absorbance of the hybridization bands in a Joyce-Loebl Chromoscan 3 densitometer as described by Pérez-Grau et al. (1986).

(d) Nucleotide sequencing

Nucleotide sequence was determined using both the dideoxy (Sanger et al., 1977) and chemical modification (Maxam and Gilbert, 1980) sequencing techniques. Internal *Pst*I, *Bam*HI, *Sac*I and *Kpn*I sites were used for subcloning the cDNA inserts in M13mp18 and M13mp19, and pUC18 and pUC19.

Sequences were determined on both strands and restriction enzyme sites used for cloning were confirmed by sequencing overlapping DNA fragments.

Whenever possible, sequence was determined by the dideoxy method using the M13 subclones. 5′- and 3′-end *Pst*I fragments (bearing the poly (dG)/(dC) tails) were sequenced using the chemical modification technique. Subclones in pUC18 were cut at the *Hin*dIII or *Bam*HI sites in the polylinker, 5′ end-labeled with polynucleotide kinase, and cut with a second restriction endonuclease. Chemical degradations were performed according to Maxam and Gilbert (1980). The A + G reaction was modified as described by Bernard and Gough (1980). Sequence alignment and treatment was done by using the computer and software of CITI2 database (Paris).

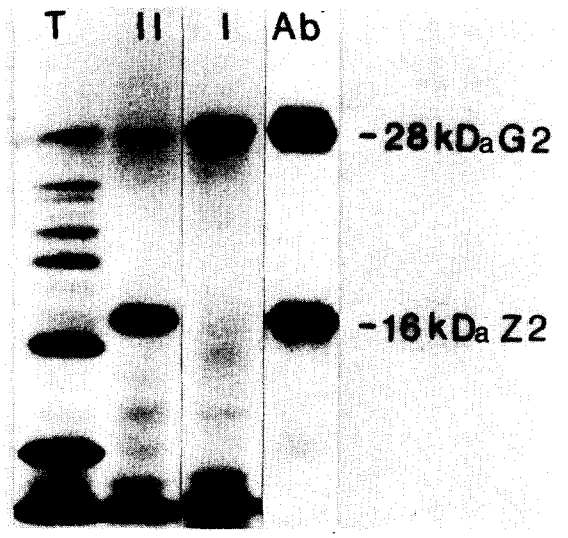


Fig. 1. In vitro translation of mRNAs selected by plasmid DNA of two subfamilies of cDNA clones from maize endosperm. Lanes: T, translation products of maize seed poly(A)$^+$ mRNA; Ab, in vitro translation products immunoprecipitated by rabbit anti-glutelin-2 antibodies; I, in vitro translation of mRNA selected by pME125 plasmid DNA; II, in vitro translation of mRNA selected by pME 177 plasmid DNA. On the right margin the relation of the two polypeptides immunoprecipitated by anti-glutelin-2 antibodies with the 28-kDa glutelin-2 and the 16-kDa zein-2 is indicated.

RESULTS AND DISCUSSION

(a) cDNA cloning and characterization

Several partial cDNA clones corresponding to a 28-kDa glutelin-2 protein were identified by screening a cDNA library from immature maize endosperm with antibodies elicited against the pure protein (Prat et al., 1985). Due to the intrinsic immune cross-reaction between the two proteins (Ludevid et al., 1985) partial clones corresponding to a 16-kDa zein-2 protein were also obtained. By using the RNase H-DNA polymerase I method for the synthesis of cDNA (Gubler and Hoffman, 1983) a new library was constructed aimed at obtaining full-length cDNA clones. This was screened with the partial cDNA clones previously obtained. Clones pME177 and pME125 with the longest inserts were chosen, they selected mRNAs which direct the synthesis 'in vitro' of 16-kDa and 28-kDa polypeptides, respectively (Fig. 1). The sequence of clone pME125 is identical with the published sequence of clone pME119 (Prat et al., 1985) except that it is longer at the 5′ end.

The length and accumulation of both mRNAs were studied by Northern analysis. Despite the difference in apparent M_r of the selected proteins, clones pME177 and pME125 hybridize to RNAs of very similar but reproducibly distinct lengths, 1180 and 1060 nt for clones of the pME125 and pME177 subfamilies, respectively (Fig. 2). The main proteins in maize endosperm, including glutelin-2, are synthesized during a well-defined stage of endosperm maturation, between 15 and 30 days after pollination (Pérez-Grau et al., 1986). This is also the observed accumulation period of mRNA encoding the 16-kDa protein as detected by Northern analysis (Fig. 2). The opaque-2 mutation that strongly reduces the level of mRNA corresponding to zeins has little effect on the accumulation of this mRNA (Fig. 2). The mRNA for this protein is located in the protein bodies fraction as are zeins and glutelin-2 (result not shown). Therefore, the biosynthesis of the proteins studied here occurs during a period and at a location typical of storage proteins and are probably regulated by the same type of signals.

(b) Amino acid sequence of the 16-kDa protein and comparison with related polypeptides

Nucleotide sequences were obtained by the M13-dideoxy method. The chemical modification proce-

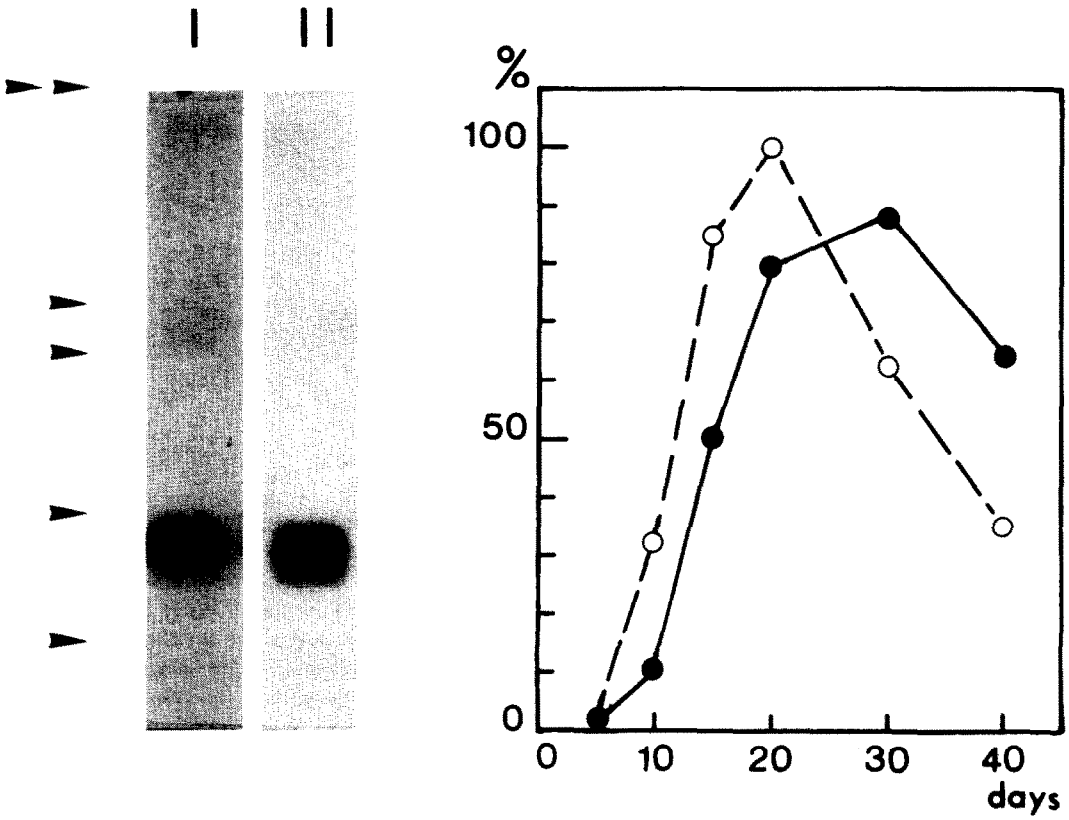
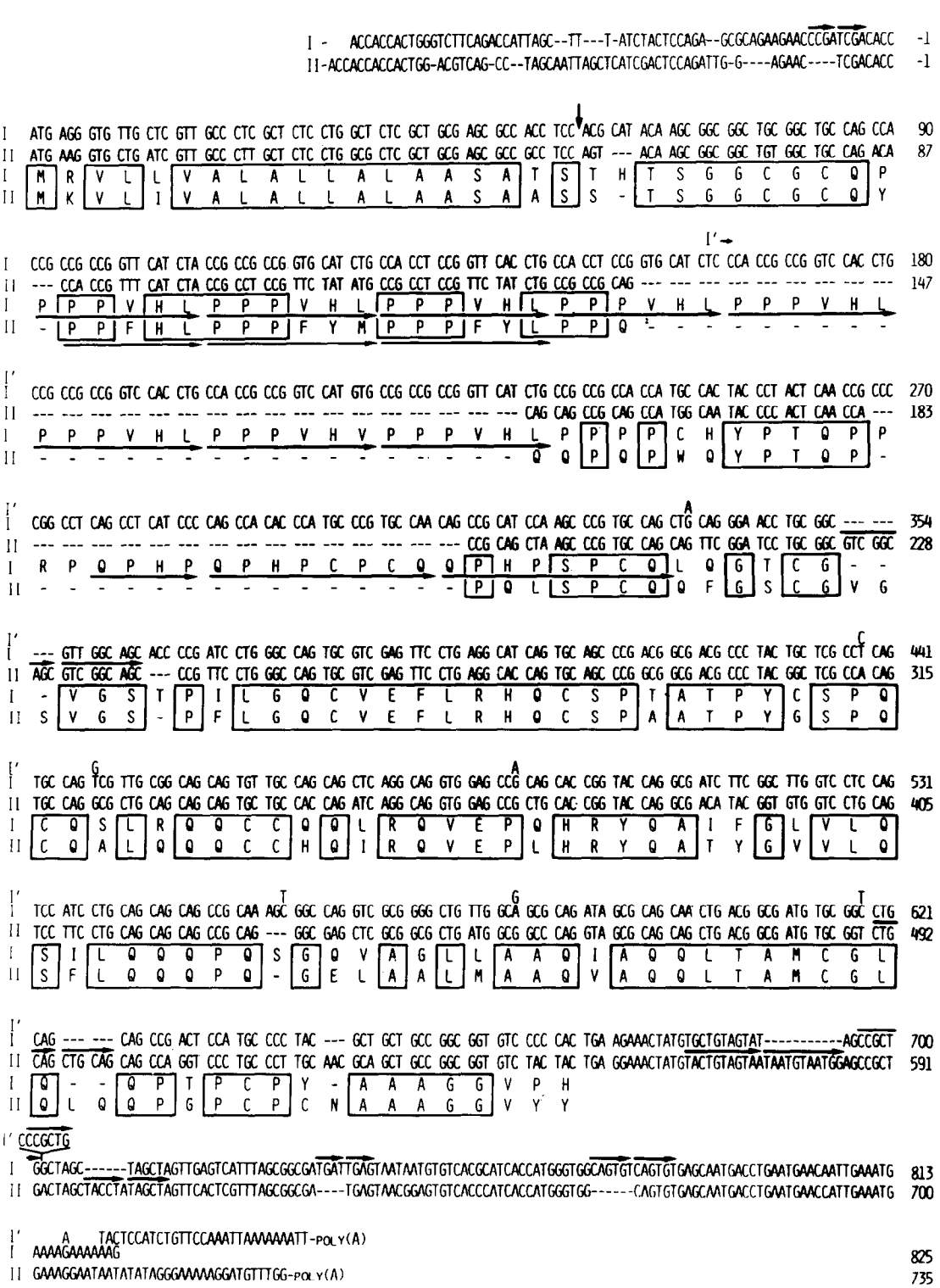


Fig. 2. Northern blot analysis of RNAs corresponding to 28-kDa glutelin-2 and 16-kDa zein-2 proteins and quantification of mRNA level during seed maturation. Lanes: I, blot hybridized with nick-translated pME125; II, blot hybridized with pME177. Black arrowheads indicate the migration of M_r markers, double arrowhead indicates the origin of the gel. Relative abundance of specific mRNAs was determined by blotting RNA extracted from immature maize seeds at different days after pollination and densitometry of the autoradiography. Black circles (solid line) represent the results for W64 inbred line; and open circles (broken line) the corresponding opaque-2 version.

dure was used to sequence through the dC/dG tails. The results are shown in Fig. 3. Sequence I corresponds to the 28-kDa glutelin-2 protein and sequence II to the 16-kDa zein-2 protein. A partial sequence of a different cDNA clone (pME792, sequence I′) belonging to the 28-kDa group but having a different restriction map is also shown for comparison. Another sequence of this group of proteins has been published (Wang and Esen, 1986) being identical to the published sequence of pME119. This identity has allowed these authors to propose that this might form a small multigene family. We have observed only three different restriction maps (result not shown) among the clones obtained corresponding to type I sequences, a result that agrees with this proposal.

The M_rs of the two mature proteins as calculated from the sequence are 21780 for the glutelin-2 protein and 17790 for the zein-2 protein. These values differ significantly from those obtained by gel electro-

Fig. 3. Comparative analysis of nucleotide and amino acid sequences from 28-kDa glutelin-2 and 16-kDa zein-2 cDNAs. Nucleotides are numbered beginning with the ATG start codon. Gaps have been introduced into the amino acid and nucleotide sequences to achieve maximum sequence homology. Conserved residues are boxed. Duplications are indicated by arrows on the nucleotide sequence. Arrows under the amino acid sequence denote position of tandem repeat units. The end of the signal peptide in the amino acid sequence is shown by a vertical arrow. Complete nucleotide sequence of 28-kDa glutelin-2 (sequence I) mRNA was deduced from clone pME125. Sequence II was deduced from clone pME177. Sequence of clone pME792 (sequence I′), belonging to the 28-kDa group, but showing a different restriction map is also included; only changes in the nucleotide sequence are shown, the protein sequence is identical except for one replacement (Thr for Ser) in position 157.



phoresis. This difference may be attributed to an unusual electrophoretic mobility, particularly for the 28-kDa protein, due to the peculiar amino acid composition of these proteins and to the structure of the repetitive domain (Tathman et al., 1985). Assuming that these clones are near to full-length, the corresponding mRNAs differ by only 127 nt, a value that agrees with that observed by Northern analysis (Fig. 2).

The deduced amino acid sequences start with a peptide absent in a partial glutelin-2 aa sequence obtained by protein sequencing (Esen et al., 1982). This fragment has the properties typical of a signal peptide and is similar in both proteins showing only three replacements, two of them conservative. In the sequence of the mature proteins a short conserved N-terminal segment containing two cysteines precedes a repetitive region that is formed by eight repetitions of the sequence Pro-Pro-Pro-Val-His-Leu peptide (with only a Leu/Val replacement) in sequence I while sequence II has only two repeats, Pro-Pro-Pro-Phe-Tyr-(Met/Leu) and an incomplete one (Pro-Pro-Phe-His-Leu). After a short (11 aa residues) common proline-rich sequence, another repetitive stretch is present in sequence I but absent in sequence II. This sequence can be considered as formed by a twice repeated octapeptide that is in turn partially repeated as a tetrapeptide (Fig. 3) giving rise to a 21-aa long Pro-X stretch. The C-terminal part of both proteins can be easily aligned by introducing small insertions or deletions. In spite of a large number of aa substitutions observed in the C-terminal part of the proteins the aa composition of these domains is not altered. The C-terminal half of these proteins contains segments having homology with storage proteins from other cereals (Prat et al., 1985). Up to now this is to our knowledge the only group of maize-seed proteins showing significant homology to the superfamily of storage proteins (Kreis et al., 1985). It is interesting to note that the sequence segments having homology with storage proteins of other cereals have also a low but significant degree of homology with a methionine-rich 14-kDa protein from maize endosperm (Marks et al., 1985) suggesting that this may belong to the same superfamily of proteins.

(c) Comparison of the nucleotide sequences

The comparison of the sequence of the cDNA clones analyzed allows to identify different types of sequence variation. One type of divergence between the two sequences arises from the generation of repetitive sequences. The presence of this type of sequences is common in cereal storage proteins (Kreis et al., 1985). In the case of the 28-kDa glutelin-2 two types of repetitive sequences are present in the N-terminal half of the protein. The lack of most of these repetitive segments in the 16-kDa protein indicates that the generation of these repeats is a phenomenon probably subsequent to a gene duplication event that generated the two sequences. The nucleotide sequence of the hexapeptide repeats is highly conserved in sequence I with only 1 to 4 nt changes from a consensus sequence (CCG CCG CCG GTC CAT CTG). In sequence II there are two repeats that vary among themselves in a single base pair, a third repeat is not complete and has a higher number of changes. The repetitive hexapeptide shows a striking similarity with repetitive sequences observed in the hydroxyproline-rich glycoprotein extensin (Chen and Varner, 1985). This homology may indicate a possible structural function for these seed proteins as has been previously proposed on the basis of their location in the inner surface of the protein bodies' membrane (Ludevid et al., 1984).

Another type of variation observed is the presence of a number of insertions distributed along both sequences. All the insertions in the coding region contain a number of nucleotides that is a multiple of three, and therefore they do not alter the reading frame of the protein. This is not the case in the 5′ and 3′ non-translated sequences. The majority of these insertions result from duplications, perfect in most of the cases, of adjacent sequences (Fig. 3). The main difference between sequences I and I′, corresponding to a protein having a single change in its amino acid sequence, is a duplicate insertion of 7 nt in the 3′ non-coding region. These duplications may be interpreted as resulting from DNA polymerase slippage during replication, however, the possibility that this type of sequence divergence can be generated by transposable elements in plants has recently been proposed (Schwarz-Sommer et al., 1985). The last type of variation observed is single-nucleotide changes that produce sequence divergence in the

different regions of the gene. In our case they are as frequent in the first (31 nt changes) as in the third (29 nt changes) codon base in the C-terminal region of the protein. However, the overall amino acid composition of this region has been conserved and in particular the number of essential amino acids has been maintained, particularly Cys and Gln residues. It has been proposed that Gln may act as the main nitrogen reservoir in seed proteins (Tsai et al., 1980) while Cys may be important for keeping the interactions between storage proteins in the outer part of the protein bodies (Ludevid et al., 1984).

(d) Evolutionary hypotheses

Four distinct types of changes in the genome can be distinguished to explain the differences between the two sequences studied: the duplication of the gene itself, the generation of repeated sequences, the insertion of sequences by duplication of small segments and single-nucleotide changes. This is summarized in Fig. 4. The four types of changes observed in the sequences studied may originate from four independent mechanisms of generation of variability. In fact, repeated sequences similar to those observed in these maize proteins have also been found in the sequence of several cereal storage proteins and in phaseolins (Slightom et al., 1985). However, in phaseolins no insertions arising from duplication of adjacent sequences are observed. This observation could indicate that these may be generated by two independent mechanisms. It is interesting to note that no intron has been observed in any cereal storage protein gene. Therefore, splicing (Gilbert, 1978) is probably not the origin of sequence variability in these storage proteins that have, nevertheless, a clear domain structure. However, it is possible that cereal seed proteins should be considered as proteins generated from a single domain (containing several conserved modules; see Kreis

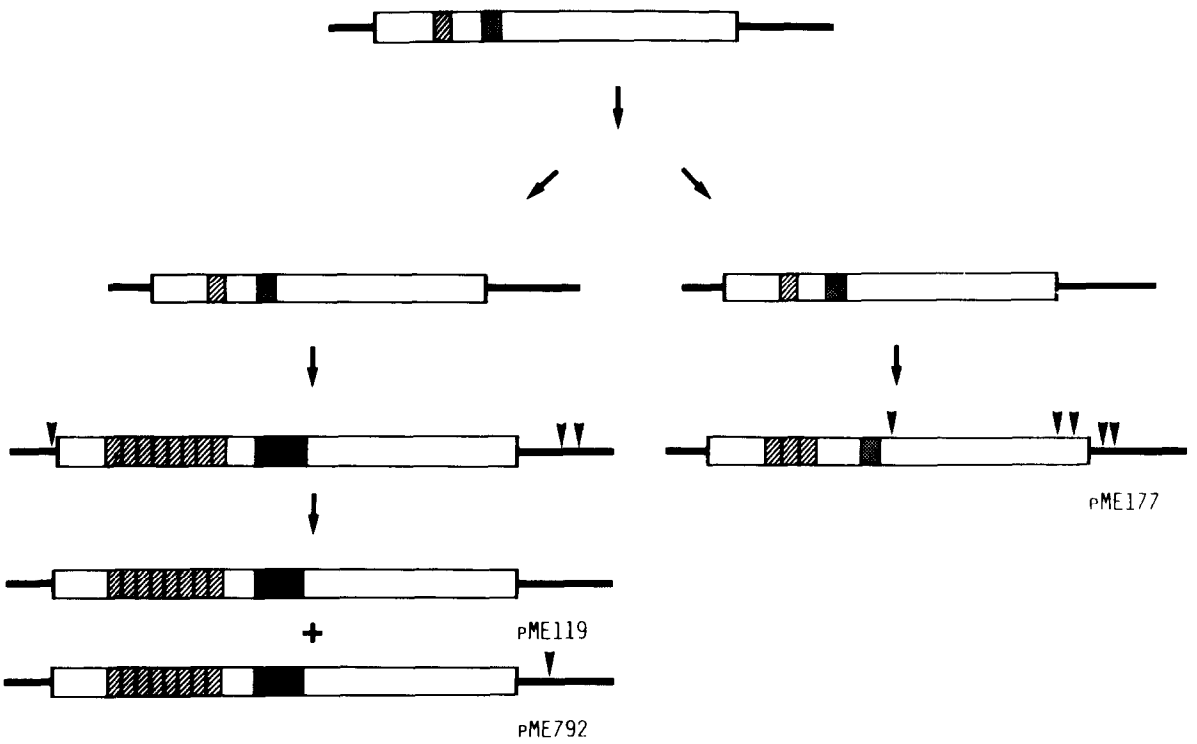


Fig. 4. Proposed evolutionary pathway of glutelin-2 and zein-2 sequences. The sequences corresponding to the three main mRNAs studied are shown. Open boxes correspond to coding and single lines to non-coding segments. Shaded and hatched boxes indicate the situation of the repetitive sequences. Downward arrowheads indicate the presence of short duplicative insertions. Clones pME119 and pME792 code for 28-kDa polypeptides (sequences I and I′, respectively) and clone pME177 (sequence II) to a 16-kDa polypeptide. Different independent evolutionary mechanisms are considered: the duplication of an ancestor gene, the generation of repetitive sequences, the production of duplicative insertions and the continuous production of single-nucleotide changes.

et al., 1985, and Reeck and Hedgcoth, 1985). The diversity in these proteins would be produced by adjoining repetitive sequences and by the action of mutational mechanisms such as transposable elements. Point mutations may tend to obscure the effects of such events in the nucleotide sequences. In the case of maize endosperm proteins it is possible that these effects can be revealed as a result from the strong selective pressure in the search for phenotypic variation to which maize has been subjected by the action of man.

(e) Conclusions

(1) The 16-kDa zein-2 protein has a sequence homologous to the 28-kDa glutelin-2, thus forming a (probably small) multigene family of proteins found in maize endosperm. The two proteins are expressed simultaneously (between 15 and 40 days after pollination) and they are not repressed in opaque-2 varieties. The difference in M_r between the two proteins probably arises from an anomalous electrophoretical mobility of the 28-kDa polypeptides due to the presence of Pro-rich repetitive sequences.

(2) This family of proteins has no homology with zeins but it clearly belongs to the same superfamily of proteins as α/β-gliadins, γ-gliadins, γ-secalins and B-hordeins, storage proteins from wheat, rye and barley.

(3) The comparison of the sequences at both the amino acid and nucleotide level allows to define distinct types of sequence variations. Accordingly, four independent types of sequence variation are proposed that may have acted upon the maize genome generating the sequence diversity observed in storage proteins.

ACKNOWLEDGEMENTS

The authors are indebted to the initial participation of Dr. Jordi Cortadas to whose memory this article is dedicated. The present work has been financed by CSIC (grant 618/500).

REFERENCES

Bernard, O. and Gough, N.M.: Nucleotide sequence of immunoglobulin heavy chain joining segments between translocated VH and μ constant region genes. Proc. Natl. Acad. Sci. USA 77 (1980) 3630–3634.

Chen, J. and Varner, J.E.: An extracellular matrix protein in plants: characterization of a genomic clone for carrot extensin. EMBO J. 4 (1985) 2145–2151.

Dretzen, G., Bellard, M., Sassone-Corsi, P. and Chambon, P.: A reliable method for the recovery of DNA fragments from agarose and acrylamide gels. Anal. Biochem. 112 (1981) 295–298.

Esen, A., Bietz, J.A., Paulis, J.W. and Wall, J.S.: Tandem repeats in the N-terminal sequence of a proline-rich protein from corn endosperm. Nature 296 (1982) 678–679.

Gilbert, W.: Why genes in pieces? Nature 271 (1978) 501.

Gubler, U. and Hoffman, B.J.: A simple and very efficient method for generating cDNA libraries. Gene 25 (1983) 263–269.

Hanahan, D.: Studies on transformation of *Escherichia coli* with plasmids. J. Mol. Biol. 166 (1983) 557–580.

Kreis, M., Shewry, P.R., Forde, B.G., Forde, J. and Miflin, B.J.: Structure and evolution of seed storage proteins and their genes with particular reference to those of wheat, barley and rye. Oxf. Surveys Plant Mol. Cell Biol. 2 (1985) 253–317.

Ludevid, M.D., Martínez-Izquierdo, J.A., Armengol, M., Torrent, M., Puigdomènech, P. and Palau, J.: Immunological relations between glutelin-2 and low molecular weight zein-2 proteins from maize (*Zea mays* L.) endosperm. Plant Sci. Lett. 41 (1985) 41–48.

Ludevid, M.D., Torrent, M., Martínez-Izquierdo, J.A., Puigdomènech, P. and Palau, J.: Subcellular localization of glutelin-2 in maize (*Zea mays* L.) endosperm. Plant Mol. Biol. 3 (1984) 227–234.

Marks, M.D., Lindell, J.S. and Larkins, B.A.: Nucleotide sequence analysis of zein mRNAs from maize endosperm. J. Biol. Chem. 260 (1985) 16451–16459.

Maxam, A.M. and Gilbert, W.: Sequencing end-labeled DNA with base-specific chemical cleavages. Methods Enzymol. 65 (1980) 499–560.

Okayama, H. and Berg, P.: High-efficiency cloning of full-length cDNA. Mol. Cell Biol. 2 (1982) 161–170.

Pérez-Grau, L., Cortadas, J., Puigdomènech, P. and Palau, J.: Accumulation and subcellular localization of glutelin-2 transcripts during maize endosperm development. FEBS Lett. 202 (1986) 145–148.

Prat, S., Cortadas, J., Puigdomènech, P. and Palau, J.: Nucleic acid (cDNA) and amino acid sequences of the maize endosperm protein glutelin-2. Nucl. Acids Res. 13 (1985) 1493–1504.

Reeck, G.R. and Hedgcoth, C.: Amino acid sequence alignment of cereal storage proteins. FEBS Lett. 180 (1985) 291–294.

Rigby, P.W.J., Dieckmann, M., Rhodes, C. and Berg, P.: Labelling deoxyribonucleic acid to high specific activity in vitro by nick translation with DNA polymerase I. J. Mol. Biol. 113 (1977) 237–251.

Sanger, F., Nicklen, S. and Coulson, A.R.: DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. USA 74 (1977) 5463–5467.

Schwarz-Sommer, Z., Gierl, A., Cuypers, H., Peterson, P.A. and Saedler, H.: Plant transposable elements generate the DNA sequence diversity needed in evolution. EMBO J. 4 (1985) 591–597.

Slightom, J.L., Drong, R.F., Klassy, R.C. and Hoffman, L.M.: Nucleotide sequences from phaseolin cDNA clones: the major storage proteins from *Phaseolus vulgaris* are encoded by two unique gene families. Nucl. Acids Res. 13 (1985) 6483–6498.

Soave, C. and Salamini, F.: Organization and regulation of zein genes in maize endosperm. Phil. Trans. R. Soc. Lond. B 304 (1984) 341–347.

Tathman, A.S., Drake, A.F. and Shewry, P.R.: A conformational study of a glutamine and proline rich cereal seed protein, C hordein. Biochem. J. 226 (1985) 557–562.

Tsai, C.Y., Huber, D.M. and Warren, H.L.: A proposed role of zein and glutelin as N sinks in maize. Plant Physiol. 66 (1980) 330–333.

Villa-Komaroff, L., Efstratiadis, A., Broome, S., Lomedico, P., Tizard, R., Naker, S.P., Chick, W.L. and Gilbert, W.: A bacterial clone synthesizing proinsulin. Proc. Natl. Acad. Sci. USA 75 (1978) 3727–3730.

Wang, S.Z. and Esen, A.: Primary structure of a proline-rich zein and its cDNA. Plant Physiol. 81 (1986) 70–74.

Wienand, U. and Feix, G.: Electrophoretic fractionation and translation in vitro of poly(A^+)-containing RNA from maize endosperm. Eur. J. Biochem. 92 (1978) 605–611.

Communicated by J-P. Lecocq.