The EMBO Journal Vol.1 No.11 pp.1337 – 1342, 1982

Primary structure of a genomic zein sequence of maize

Nien-Tai Hu, Mark A. Peifer[1], Gisela Heidecker[3], Joachim Messing[2], and Irwin Rubenstein*

Department of Genetics and Cell Biology, and [2]Department of Biochemistry, University of Minnesota, St. Paul, MN 55108, [1]Department of Biochemistry and Molecular Biology, Harvard University, Cambridge, MA 02138, and [3]Department of Human Pathology, School of Medicine, University of California, Davis, CA 95616, USA

Communicated by P. Starlinger
Received on 1 September 1982

The nucleotide sequence of a genomic clone (termed Z4) of the zein multigene family was compared to the nucleotide sequence of related cDNA clones of zein mRNAs. A tandem duplication of a 96-bp sequence is found in the genomic clone that is not present in the related cDNA clones. When the duplication is disregarded, the nucleotide sequence homology between Z4 and its related cDNAs was ~97%. The nucleotide sequence is also compared to other isolated cDNAs. No introns in the coding region of the zein gene are detected. The first nucleotide of a putative TATA box, TATAAATA, was located 88 nucleotides upstream of the first nucleotide of the first ATG codon which initiated the open reading frame. The first nucleotide of a putative CCAAT box, CAAAAT, appeared 45 nucleotides upstream of the first nucleotide of the TATA box. The possible polyadenylation signals found in the zein cDNA clones in the 3′ non-coding region also appeared in the genomic sequence at the same locations. The amino acid composition of the polypeptide specified by the Z4 nucleotide sequence is similar to the known composition of zein proteins.

Key words: cDNA cloning/introns/multigene/polyadenylation signals/TATA box

Introduction

The zein proteins consist of a large family of hydrophobic proteins which constitute >50% of the protein in maize kernels. They are rich in glutamine, leucine, proline, and alanine, low in arginine, histidine, and very low in lysine and tryptophan residues (Wall, 1964). On SDS-polyacrylamide gels, the proteins are resolved into two predominant bands with apparent mol. wts. of 19 000 and 22 500 (Gianazza *et al.*, 1976). Isoelectric focusing (Righetti *et al.*, 1977) and two-dimensional gel electrophoresis (Hagen and Rubenstein, 1980) reveal more complex protein patterns.

This complexity is reflected in the zein mRNAs. Different groups of zein mRNA templates representing a given subfamily are selected from the zein mRNA population by their homology to cDNA clones of individual zein mRNAs (Burr *et al.*, 1982). When each of these groups of mRNAs termed a subfamily are translated *in vitro* and analyzed on an isoelectric focusing gel, a distinctive polypeptide pattern is seen (Park *et al.*, 1980). Three zein subfamilies (labelled A20, A30, and B49 after the names of the cDNAs) have been identified by these and other experiments (Hagen and Rubenstein, 1980; Rubenstein, 1982).

Maize nuclear DNA segments isolated from the leaf tissue of the inbred W22 were incorporated into lambda bacteriophage particles to form a partial genomic library (Maniatis *et al.*, 1978; Lewis *et al.*, 1981). The cDNA clone A30 was used as a probe to find and characterize phage particles which contain homologous genomic segments. One of these phage particles was selected for further study and named λ(W22)Z4 or Z4 for short. The Z4 and A30 DNAs hybridize to a similar subfamily of mRNAs (Lewis *et al.*, 1981).

In this study, we subcloned the region of λ(W22)Z4 that hybridized to the A30 sequence and sequenced the coding and flanking regions *via* the M13 subcloning/dideoxy sequencing method. We also prepared and isolated several cDNA clones from W22 mRNAs that are homologous to Z4 and sequenced one of them. The nucleotide sequences demonstrate that the genomic clone and the two cDNA clones are closely related and represent members of the same zein subfamily (Rubenstein, 1982). We did not find any introns in the zein genomic sequence, although we did locate a putative TATA box (Efstratiadis *et al.*, 1980), a putative CAT box (Efstratiadis *et al.*, 1980), and two putative polyadenylation signals (Benoist *et al.*, 1980; Fitzgerald and Shenk, 1981) in the flanking sequences, suggesting that the Z4 clone may represent a functional zein gene.

Results and Discussion

Nucleotide sequences

The nucleotide sequence of the genomic clone, λ(W22)Z4, was determined and the message strand is shown in Figure 1. In addition, the nucleotide sequences of the related cDNA clone A30 (Geraghty *et al.*, 1981), which was prepared from IHP maize mRNA, and the cDNA clone ZG31A, which was prepared from W22 maize mRNA are also shown. The nucleotide sequence of Z4 includes 804 nucleotides of the apparent coding sequences of a zein protein and 560 nucleotides of non-translated sequence, of which 458 residues are at the 5′ terminus and 102 are at the 3′ terminus.

A comparison of the genomic and cDNA sequences demonstrates that no intron-like sequences are present in the genomic clone. Wienand *et al.* (1981) did not detect any introns in their genomic clone by electron microscopic examination of R-loops formed between the clone and endosperm poly(A) RNA. They, however, could not have seen introns shorter than 50 bp. Introns have been found in the phaseolin gene of French bean (Sun *et al.*, 1981) and the leghaemoglobin gene of soybean (Jensen *et al.*, 1981; Hyldig-Nielsen *et al.*, 1982).

A major tandem duplication is revealed, however, when the nucleotide sequence of the Z4 genomic clone is compared to the nucleotide sequences of the cDNA clones A30 and ZG31A. Depending on where one envisions the start of the duplication to occur, there are one or two silent base substitution differences between the duplicated sequences (Figure 1 and Table I). These differences suggest that the duplication did not occur during the bacterial replication of the cloned DNA. Furthermore, these additional nucleotides code for a unique region in the amino acid sequence of the encoded zein

*To whom reprint requests should be sent.

protein; the duplication occurs in the region of the zein polypeptide where two neighboring prolines are separated by the largest number of residues. Since the proline residues may play an important role in the folding of the protein, this finding suggests that the addition of 32 amino acids at this point in the polypeptide sequence may be structurally significant.

Sequences conserved at the splicing junctions of other systems, (exon)/GT------AG/(exon) (Lerner *et al.*, 1980; Rogers and Wall, 1980), also occur near the junctions of the duplication (Figure 1). Nevertheless, the removal at these sites of

```
                               CGAGTGAT TCTTTAAACC GATTATTACA CAAGTTAACC ACACTAAAAT TAACATTGGT
       GAATCGTGCC ATGATTTTTT TCTAGTGCAA AATAGCCAAA CCAAGCAAAA CATATGTGGC TATCGTTACA CATGTGTAAA
       GGTATTGCAT CACACCATTG TCACCCATGT ATTTGGACAA TACCGAGAGG AAAAACCACT TATTTATTGT ATTTTATCAA
       GTTTATCTTG CTTACGTATA AATTATAACC CAACAAAGTA ATCACTAAAT GTCAAAACCA ACTAGATACC ATGTCATCTC
       TACCTTATCT TACTAATATT CTTTTTGCAA AATCGAAAAT TAATCTTGCA CAAGCACAAG GACTGAGATG TGTATAAATA
       TCTCTTAGAT TAGTAGATAA TATATCGCAC ATATTATTGA GACCAACTAG CAACATAGAA AGCACAATAT TGTACCAATA
                                                                                      A30... C
                                                             10
Z4     ATG   GCA   GCC   AAA   ATA   TTT   TGC   CTC   ATT   ATG   CTC   CTT   GGT   CTT   TCT   GCA   AGT
A30                                                     C
                   20                                                            30
Z4     GCT   GCT   ACG   GCG   AGC   ATT   TTC   CCG   CAA   TGC   TCA   CAA   GCT   CCT   ATA   GCT   TCC
A30                             C                                   A
ZG31A...............CG          C                                   G
                                     40                                                          50
Z4     CTT   CTT   CCC   CCA   TAC   CTC   TCA   CCA   GCG   ATG   TCT   TCA   GTA   TGT   GAA   AAT   CCA
A30                       G                                   G           G                       C
ZG31A                     G                                   G           G                       C
                                                       60
Z4     ATT   CTT   CTA   CCC   TAC   AGG   ATC   CAA   CAG   GCA   ATC   GCA   GCA   GGC   ATC   TTA   CCT
A30                 A                                                     G     T
ZG31A               A                                                     A     T
             70                                                          80
Z4     TTA   TCA   CCC   TTG   TTC   CTC   CAA   CAA   TCA   TCA   GCC   CTA   TTA * CAG   CAG + TTA   CCT
A30                                                                                   G
ZG31A                                                                                 T
                               90                                                          100
Z4     TTG   GTG   CAT   TTA   TTG   GCA   CAA   AAC   ATC   AGG   GCA   CAA   CAA   CTA   CAA   CAA   CTC
A30                                                                                                     T
ZG31A                                                                                                   T
                                                 110
Z4     GTG   CTA   GCA   AAC   CTT   GCT   GCC   TAC   TCT   CAG   CAA * CAG   CAG + TTA   CCT   TTG   GTG
A30                                                                    (.......................................
ZG31A                                                                  (.......................................
       120                                                         130
Z4     CAT   TTG   TTG   GCA   CAA   AAC   ATC   AGG   GCA   CAA   CAA   CTA   CAA   CAA   CTC   GTG   CTA
A30    ....................................................................................................
ZG31A  ....................................................................................................
                         140                                                         150
Z4     GCA   AAC   CTT   GCT   GCC   TAC   TCT   CAG   CAA * CAA   CAG + TTT   CTG   CCA   TTC   AAC   CAA
A30    ...................................................)   G                 T
ZG31A  ...................................................)   G                 T
                                           160                                                         170
Z4     CTA   GCT   GCA   TTG   AAC   TCT   GCT   GCT   TAT   TTG   CAG   CAA   CAA   CAA   CTA   CTA   CCA
A30                                               T                 A                     (...)
ZG31A                                             T                 A                     (...)
                                                             180
Z4     TTC   AGC   CAG   CTA   GCT   GCT   GCC   TAC   CCC   CGG   CAA   TTT   CTT   CCA   TTC   AAC   CAA
A30                             C                             A
ZG31A                           C                             A
                   190                                                           200
Z4     CTG   GCA   GCA   TTG   AAC   TCT   CAT   GCT   TAT   GTA   CAA   CAA   CAA   CAA   CTA   CTA   CCA
A30                                         C                 T     G     G
ZG31A                                       C                 T     G     G
                                     210                                                         220
Z4     TTC   AGC   CAG   CTA   GCT   GCT   GTG   AGC   CCT   GCT   GCC   TTC   TTG   ACA   CAG   CAA   CAT
A30                                   G                             A                       A     C     G
ZG31A                                 G                             A                       A     C     G
                                                       230
Z4     TTG   TTG   CCG   TTC   TAC   CTG   CAC   ACT   GCG   CCT   AAC   GTT   GGC   ACC   CTC   TTA   CAA
A30                                   A           G                       C
ZG31A                                 A           G                       C
             240                                                         250
Z4     CTG   CAA   CAA   TTG   CTG   CCA   TTC   GAC   CAA   CTT   GCT   TTG   ACA   AAC   CCA   GCA   GTG
A30                                               A                                         T           C
ZG31A                                             A                                         C           C
                               260
Z4     TTC   TAC   CAA   CAA   CCC   ATC   ATT   GGT   GGT   GCC   CTC   TTT   TAG   ATTGCTTATG AGTTATAGTT
A30                                                                                      T
ZG31A                                                                                    T

Z4     CAATAATAAA GTTTTTTTTG CTGATATTTG TGGCTTCCCA GAAATAAGAA AGTACATTTC TAGATTCTTA TGTGCTTCTA GT
A30               (.)     GT      G T                                        TTCT(POLY A)
ZG31A             (.)     GT      G G                                        ....(POLY A)
```

Fig. 1. The complete nucleotide sequence of a zein gene, Z4, and its flanking regions. The sequences of two cDNA clones are also included. Only the nucleotides of the cDNAs that differ from the Z4 sequence are shown. A30 was prepared from IHP maize mRNA. ZG31A was prepared from W22 maize mRNA. The Z4 gene was isolated from W22 leaf DNA. The A30 sequence starts 2 bp upstream of the first ATG. The ZG31A sequence starts 59 bp downstream of the first ATG. .... Denotes the sequences in Z4 that are not present in the cDNA clones. The * marks the start and end of the 96-bp duplications viewed one way, and the + marks the duplications viewed a different way. Codon no. 89 (TTA) and codon no. 121 (TTG) differ by a single base. Both code for leucine. Codon no. 114 (CAG) and codon no. 146 (CAA) code for the same amino acid, glutamine. The putative CAT box (CAAAT) and TATA box (TATAAATA), the first two ATG codons, the first terminating TAG, and the putative polyadenylation signals (AATAAA and AATAAG) are underlined. Also underlined are the possible splicing sites near the junctions of the 96-bp duplications and the consequent premature termination codon. The coding region is numbered every tenth codon, starting from the first ATG that is in the reading frame, □ marks the first amino acid codon of the mature zein protein as suggested by Bietz *et al.* (1979). The complete coding region of the genomic clone Z4 and 85% of the cDNA clone ZG31A were sequenced in both orientations. The non-coding regions of Z4 are sequenced only in one orientation.

Table I. Differences in amino acids among the three clones Z4, A30, and ZG31A as a result of their differences in nucleotide sequences

Codon number[a]	Amino acid encoded in clones		
	Z4	A30	ZG31A
9	Ile	Leu	[b]
22	Ser	Thr	Thr
44	Met	Val	Val
54	Leu	Gln	Gln
63	Ala	Ala	Thr
82	Gln	Gln	His
161	Ala	Ser	Ala
168	Leu	abs	abs
175	Ala	Pro	Pro
180	Arg	Gln	Gln
194	His	Pro	Pro
197	Val	Leu	Leu
210	Ala	Gly	Gly
215	Ala	Thr	Thr
220	Gln	Pro	Pro
221	His	Gln	Gln
227	Leu	Gln	Gln
229	Thr	Ala	Ala
233	Val	Ala	Ala
246	Asp	Asn	Asn
253	Pro	Leu	Pro
255	Val	Ala	Ala
114–145	32 amino acids[c]	abs	abs

[a] As designated in Figure 1.
[b] ZG31A clone does not include this region.
[c] Codons 114–145 encode for the same amino acids encoded by codons 82–113.
abs = absent.

either or both repeats would result in an early termination of protein synthesis (Figure 1). A short addition of a triplet to the Z4 sequence relative to the cDNAs is also present (Figure 1). The authors realize that these regions of apparent additions to the Z4 sequence could be also considered as regions of deletion in the cDNAs. For ease of discussion we have chosen to describe them as regions of duplication or insertion in the genomic sequence.

The nucleotide sequence of the cDNA clone ZG31A, prepared from W22 maize mRNA, is ~99% homologous to the nucleotide sequence of the cDNA clone A30 that was prepared from IHP maize (mRNA) (Figure 1). Their amino acid sequences show ~98% homology (Tables I and II). Both the 32 amino acid and the single amino acid duplications seen in the Z4 genomic sequence are absent in these cDNA clones (Figure 1 and Table I). When the duplicated regions are not included, there is a 97% homology between the Z4 and A30 (and ZG31A) nucleotide sequences.

Of 30 (31 in the case of ZG31A) single base differences in the coding region, 11 (13 in the case of ZG31A) are silent (Figure 1 and Table I). Over 80% of the differences are located in the two-thirds of the coding region at the carboxyl terminus, where tandem repetitions of the amino acid sequence have been previously reported (Table I) (Geraghty *et al.*, 1981). Lysine and tryptophan are not found in the polypeptides specified by the Z4, A30, or ZG31A nucleotide sequences (Table II).

Two ATG codons appear at the 5′ terminus of the open reading frames of the Z4 and the A30 sequences (Figure 1). A corrected version of the A30 sequence is shown in Figure 1. When the previously published sequence for the A30 cDNA (Geraghty *et al.*, 1981) was reexamined, an extra C base was found to have been erroneously included near the 5′ end of the published sequence for A30 (unpublished data). The cDNA clone, ZG31A, prepared from W22 maize mRNA, is not as long a copy of its zein mRNA as is A30; we do not know if two ATG codons are also present at the 5′ end of the ZG31A sequence. The AUG codon most proximal to the 5′ terminus is thought to be the initiation codon in yeast mRNAs (Stewart *et al.*, 1971; Kozak and Shatkin, 1978). It may be necessary to obtain amino acid sequences of the zein protein precursors to determine which of the AUGs in these zein mRNAs is used to initiate protein synthesis.

Molecular weight of zein proteins

After the signal peptide has been subtracted, the mol. wt. of the polypeptide encoded by the Z4 genomic sequence is 27 000 [Table II; see Geraghty *et al.* (1981) for a discussion of how the amino terminus of the zein protein is defined]. We assume that the zein mRNA is not spliced. On the other hand, A30 encodes for a mature protein of 23 300 (Geraghty *et al.*, 1981), and was suggested to represent a message coding for the smaller mol. wt. class of zein proteins ('19 kd'). Although both the Z4 and A30 nucleotide sequences primarily hybrid-select mRNAs coding for the smaller zein proteins, larger zein proteins ('22.5 kd') do appear on the autoradiograms after longer exposure (Lewis *et al.*, 1981). If the Z4 gene is transcribed into an mRNA and translated without being processed, it is likely to belong to the zein mRNAs coding for the larger mol. wt. class of zein. The fact that Z4 hybrid-selects predominantly the smaller zein mRNA suggests that probably a large proportion of the transcribed zein genes that are homologous to Z4 will be found not to contain the 96-bp duplication. Alternatively, genes that do not contain the duplication might be transcribed/translated with a greater frequency than the ones that contain it.

Flanking regions

An octanucleotide, TATAAATA (the first T starting at a position 88 nucleotides upstream of the first nucleotide of the first ATG) is present in the 5′-flanking sequence of the Z4 genomic clone (Figures 1 and 2). This sequence is similar to the TATA boxes reported for other eukaryotic genes (Figures 1 and 2) (Efstratiadis *et al.*, 1980). A comparison of the Z4 genomic sequence with a longer (possibly complete) cDNA sequence (A20, Geraghty *et al.*, 1982) suggests that this octanucleotide is ~30 residues upstream of the 5′ end of the mRNA. This is similar to the location found in other eukaryotes (Efstratiadis *et al.*, 1980). The assignment of the actual 5′ terminus of the mRNA will have to be confirmed by S1 mapping (Berk and Sharp, 1977) and/or primer extension experiments (Bina-Stein *et al.*, 1979). A putative CCAAT sequence (Efstratiadis *et al.*, 1980), CAAAAT, is also found upstream of the TATA box, with 45 residues separating the first nucleotides of the two sequences (Figures 1 and 2). Again this distance is about the same as has been observed in other eukaryotic gene sequences (Efstratiadis *et al.*, 1980).

The 3′-flanking sequence of the Z4 genomic sequence contains the putative polyadenylation signals, AATAAA and AATAAG. These signals occur in the A30 sequence (Geraghty *et al.*, 1981), and are also found in the ZG31A se-

Table II. Comparison of amino acid compositions

Residue	Zein clones Number of amino acids[a] and (g amino acid/100 g protein)			Zein protein[b] g amino acid/100 g protein	
	Z4	A30	ZG31A	'19 000 d'	'22 500 d'
Nonpolar					
Ala	35 (9.2)	29 (8.8)	28 (8.5)	8.8	7.7
Ile	10 (4.2)	9 (4.4)	9 (4.4)	3.5	3.5
Leu	52 (21.7)	43 (20.9)	43 (20.8)	18.5	14.5
Met	1 (0.5)	0 (0.0)	0 (0.0)	0.4	1.9
Phe	13 (7.1)	13 (8.2)	13 (8.2)	7.4	6.7
Pro	22 (7.9)	23 (9.6)	23 (9.6)	11.2	13.7
Trp	0 (0.0)	0 (0.0)	0 (0.0)	n.d.	n.d.
Val	9 (3.3)	5 (2.1)	5 (2.1)	3.4	4.2
Sum	142	122	121		
Polar					
Asn	11 (4.7)	10 (4.9)	10 (4.9)	5.0	4.1
Cys	2 (0.8)	2 (0.9)	2 (0.9)	–	–
Gln	47 (22.2)	41 (22.5)	40 (21.9)	21.2[c]	18.0[c]
Gly	4 (0.9)	5 (1.2)	5 (1.2)	2.1	4.2
Ser	16 (5.1)	15 (5.6)	15 (5.6)	5.3	4.5
Thr	4 (1.8)	5 (2.2)	6 (2.6)	2.5	2.4
Tyr	9 (8.4)	8 (5.6)	8 (5.6)	4.8	6.4
Sum	93	86	86		
Basic					
Arg	4 (2.3)	2 (1.3)	2 (1.3)	2.4	4.1
His	5 (2.5)	2 (1.2)	3 (1.8)	1.2	2.2
Lys	0 (0.0)	0 (0.0)	0 (0.0)	–	–
Sum	9	4	5		
Acidic					
Asp	1 (0.4)	0 (0.0)	0 (0.0)	c	c
Glu	1 (0.5)	1 (0.6)	1 (0.6)	c	c
Sum	2	1	1		
Total number of residues	246	213	213		
Calculated mol. wt.	27 136	23 329	23 371		

[a]Signal peptide is not included.
[b]Lee *et al.*, 1976.
[c]We assumed that the aspartic acid and glutamic acid residues were in the form of asparagine and glutamine.

quence (Figures 1 and 2). A difference of four nucleotides exists between the ZG31A and A30 sequences near the site where poly(A) is added to the mRNA. The nucleotide sequence at the 3′ end of the genomic clone Z4, however, is identical to A30 (Figure 1).

The genomic clone Z4 could represent a functional or nonfunctional gene or a pseudogene. We have not as yet sequenced a zein cDNA clone that contains the exact sequence seen in Z4. This would be one way to prove that this sequence is functional. In systems less complex than the zein multigene family, S1 mapping of mRNA/DNA hybrids has been used to determine whether or not an mRNA was actually produced by a cloned genomic sequence (Berk and Sharp, 1977). This experiment is probably impossible to perform correctly in a system with the characteristics of the zein multigene family. The members of a subfamily are likely to differ by only a few nucleotides from one another (Figure 1). Therefore, a finding that mRNAs exist that can be protected from S1 digestion by the Z4 sequence or that can protect the Z4 sequence would not prove that the Z4 sequence was a functional gene since proper controls are not available to ensure the complete digestion of the short single-stranded regions that would be expected to occur (Wells *et al.*, 1980).

Could the genomic clone Z4 be a pseudogene? The Z4 sequence does not contain the structural characteristics such as termination codons or frameshifts that are often present in pseudogenes (Proudfoot, 1980). Moreover, the other features of the Z4 sequence mentioned above, would suggest that Z4 has the known sequences required for the transcription and processing of its mRNA. Therefore, we conclude that it probably represents a functional zein gene.

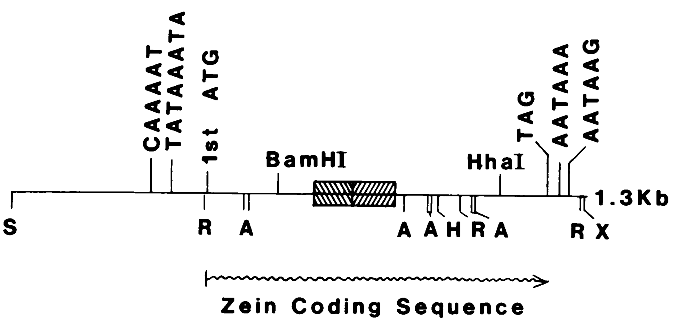

Fig. 2. Summary of the features observed in the nucleotide sequence of the genomic clone Z4. The message strand is shown with the direction of transcription from left to right. Restriction sites are designated as follows: S, *Sau*3A; R, *Rsa*I; A, *Alu*I; H, *Hpa*II; and X, *Xba*I. The hatched areas represent the duplications of a fragment of 96 nucleotides that is present once in the cDNA clone A30. The zein protein coding sequence, represented by ⇝, starts from the first ATG codon which is in the reading frame and ends at the first TAG codon. The coding strand downstream of the *Bam*HI site down to the first duplicated sequence was also sequenced by the Maxam and Gilbert method (Maxam and Gilbert, 1977). The results agree with the sequence determined by the dideoxy method.

Evolution of zein

Various repetitions are observed in the amino acid sequence of the proteins of the zein multigene family (Figures 1 and 2; Geraghty *et al.*, 1981). None of these are exact repeats. Amino acid changes, which include substitutions and deletions/insertions, occur from one repeat to another. The significance of this repetitious structure is not known. Being the major storage proteins stored in the protein body of the endosperm of maize, zein provides the nitrogen and amino acid sources for seed germination. During evolution, a certain repeating pattern may have been conserved to facilitate its maximal packaging or the most efficient degradation during the germination of the seed. Studies of the secondary structure of zein protein may clarify the role of the observed repetitions. The sequence of additional zein genes will serve to define the structural parameters of this diverse multigene family.

The subfamilies of the zein multigene family

The zein multigene family can be divided into at least three families based on the nucleotide sequence relatedness of their mRNAs (Park *et al.*, 1980). For the purpose of clarity, we have renamed these families as subfamilies. We have shown that a close sequence homology of >95% exists among the genomic clone Z4 and the cDNA clones A30 and ZG31A. They therefore belong to the A30 subfamily [named after the cDNA clone that first identified its existence (Park *et al.*, 1980)]. Assuming that the Z4 genomic sequence is transcribed *in vivo* and the mRNA is translated without being processed, then the difference in mol. wt. of the proteins that would be coded for by Z4 and A30 is equal to the difference between the larger and smaller zein proteins. This suggests that the nucleotide sequence homology observed among members of the same subfamily is not limited to zein proteins of the same size class.

Materials and methods

Materials and reagents

The restriction endonucleases *Pst*I, *Hin*dIII, *Bam*HI, *Hin*cII, *Eco*RI, *Xba*I, *Sal*I, and *Escherichia coli* DNA polymerase, large fragment (Klenow enzyme) were obtained from Bethesda Research Labs. The restriction endonucleases *Acc*I, *Alu*I, *Rsa*I, *Hpa*II, and *Sau*3A were obtained from New England Biolabs. T4 DNA ligase was from both New England Biolabs and Bethesda Research Labs. The condition for obtaining *Eco*RI* cleavage was as described (Gardner *et al.*, 1981). The terminal deoxyribonucleotide transferase was from Ratliff Biochemicals and the RNasin from Biotec., Inc. The $[\alpha\text{-}^{32}P]$dATP was from either Amersham or from New England Nuclear. The synthetic oligonucleotide of 15 residues, 5′-TCCCAGTCACGACGT-3′, which serves as the primer (Messing *et al.*, 1981) in the dideoxy sequencing reactions (Sanger *et al.*, 1977) was from New England Biolabs.

Preparation of zein cDNA clones from W22 zein mRNAs

Zein mRNAs were isolated from protein bodies (Burr and Burr, 1976) and purified using a dimethyl sulfoxide sucrose gradient (Burr *et al.*, 1978). 1 μg of mRNA was hybridized to 1 μg pUC9 (Vieira and Messing, 1982) that had been cut at the *Pst*I site and extended with thymidine triphosphate using terminal deoxynucleotide transferase (Deng and Wu, 1981). The average poly(dT) tail length was 40 nucleotides. The cDNA synthesis was then carried out for 1 h at 37°C in a final volume of 15 μl (70 mM KCl, 50 mM Tris pH 8.2, 10 mM $MgCl_2$, 1 mM dithiothreitol, 0.5 mM of each of the four deoxyribonucleoside triphosphates, 25 μg/ml actinomycin D, 20 units RNasin, and 8 units reverse transcriptase, which was kindly provided by Dr. J. Beard). The reaction was terminated by phenol extraction and the unincorporated deoxyribonucleoside triphosphates were removed by several ethanol precipitations. The conjugated cDNA was then extended with deoxyguanidine triphosphate (Deng and Wu, 1981). The mRNA and other small molecules were removed on a 5 – 20% alkaline sucrose gradient. The fractions containing single-stranded molecules longer than pUC9 were pooled and mixed with 10 μg of C-tailed pUC9. The mixture was subsequently dialyzed against 10 mM Tris, 1 mM EDTA pH 8.0 to remove the sodium hydroxide and sucrose. The molecules were renatured for 24 h at 37°C in a final volume of 3 ml of 32% formamide, 54 mM NaCl, 10 mM Tris pH 8.5 (Fanning *et al.*, 1976). Renaturation was >90% as judged by agarose gel electrophoresis of the products. The DNA was concentrated by ethanol precipitation in the presence of carrier RNA, resuspended and the second strand of the cDNA was filled in using the large fragment of *E. coli* DNA polymerase I. Transformation with the equivalent of 20 ng of starting T-tailed pUC9 gave 400 clones, 40% of which hybridized to Z4 and 60% of which had inserts >400 bp. Of the clones with zein inserts, 20% were >800 bp.

Preparation of the zein genomic fragment

A 3.6-kb *Hin*dIII fragment of the zein genomic clone λ(W22)Z4 (Lewis *et al.*, 1981), was purified and subcloned into pBR322 at the *Hin*dIII site, and then transformed into *E. coli* LE392 (thy A, sup E, sup F, mk+ rk –). The appropriate recombinant plasmid, selected by its homology to λ(W22)Z4 DNA, was amplified and isolated (Park *et al.*, 1980). The 3.6-kb *Hin*dIII fragment was cleaved from the recombinant plasmid DNA by use of *Hin*dIII and purified from an agarose gel.

M13 subcloning and dideoxy sequencing

Subfragments of the 3.6-kb *Hin*dIII fragment and those of the cDNA insert of the cDNA clone ZG31A were generated by various restriction endonucleases and cloned into the multipurpose cloning vector M13mp7 (Messing *et al.*, 1981) or force cloned into the asymmetric cloning vectors M13mp8, M13mp9, and M13mp13 (Messing and Vieira, 1982). The subclones were selected at random and by their complementarity to single strand-specific M13 probes (Hu and Messing, 1982). The double-stranded form of the phage DNAs was prepared from phage-infected cells; the recombinant M13 phage were reintroduced into *E. coli* JM103; and the single-stranded DNAs to be used as templates were prepared from the phage particles (Heidecker *et al.*, 1980). Sequences were determined by Sanger's dideoxy terminator method (Sanger *et al.*, 1977), using a synthetic primer that is complementary to the region 3′ to the multiple cloning sites (Messing *et al.*, 1981).

Computer processing of the sequence data

All of the sequence data were entered, stored on computer diskettes, and analyzed on Apple II plus computer system using the computer program developed by Larson and Messing (1982).

Acknowledgements

We thank Kris Kohn, Kae Ebling, and Yvonne Schack for their help in preparing this manuscript. This research was supported by grants from the National Institutes of Health, GM24756; from the USDA/SEA Competitive Grant Program – Genetic Mechanisms for Crop Improvement, 59-2271-0-1401-0; from the Department of Energy, DE-AC02-81ER 10901; from the Minnesota Agricultural Experiment Station, MN-15-030; and from the National Institutes of Health, ST32 GM07467-05.

References

Benoist,C., O'Hare,K., Breathnach,R., and Chambon,P. (1980) *Nucleic Acids Res.*, **8**, 127-142.

Berk,A.J., and Sharp,P.A. (1977) *Cell,* **12**, 721-732.

Bietz,J.A., Paulis,J.W., and Wall,J.S. (1979) *Cereal Chem.,* **56**, 327-332.

Bina-Stein,M., Thoren,M., Salzman,N., and Thompson,J.A. (1979) *Proc. Natl. Acad. Sci. USA,* **76**, 731-735.

Burr,B., Burr,F.A., Rubenstein,I., and Simon,M.N. (1978) *Proc. Natl. Acad. Sci. USA,* **75**, 696-700.

Burr,B., and Burr,F.A. (1976) *Proc. Natl. Acad. Sci. USA,* **73**, 515-519.

Burr,B., Burr,F.A., St.John,T.P., Thomas,M., and Davis,R.W. (1982) *J. Mol. Biol.,* **154**, 33-49.

Deng,G., and Wu,R. (1981) *Nucleic Acids Res.,* **9**, 4173-4188.

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.

Fanning,T.G., Schreier,P.F., and Davies,R.W. (1976) *Eur. J. Biochem.,* **62**, 173-179.

Fitzgerald,M., and Shenk,T. (1981) *Cell,* **24**, 251-260.

Gardner,R.C., Howarth,A.J., Hahn,P., Brown-Luedi,M., Shepherd,R.J., and Messing,J. (1981) *Nucleic Acids Res.,* **9**, 2871-2888.

Geraghty,D., Peifer,M.A., Rubenstein,I., and Messing,J. (1981) *Nucleic Acids Res.,* **9**, 5163-5174.

Geraghty,D., Messing,J., and Rubenstein,I. (1982) *EMBO J.,* **1**, 1329-1335.

Gianazza,E., Righetti,P.G., Pioli,F., Galante,E., and Soave,C. (1976) *Maydica,* **21**, 1-17.

Hagen,G., and Rubenstein,I. (1980) *Plant Sci. Lett.,* **19**, 217-223.

Heidecker,G., Messing,J., and Gronenborn,B. (1980) *Gene,* **10**, 69-73.

Hu,N.T., and Messing,J. (1982) *Gene,* **17**, 271-277.

Hyldig-Nielsen,J.J., Jensen,E.Ø., Paludan,K., Wiborg,O., Garrett,R., Jørgensen,P., and Marcker,K.A. (1982) *Nucleic Acids Res.,* **10**, 689-701.

Jensen,E.Ø., Paludan,K., Hyldig-Nielsen,J.J., Jørgensen,P., and Marcker, K.A. (1981) *Nature,* **291**, 677-679.

Kozak,M., and Shatkin,A.J. (1978) *Cell,* **13**, 201-212.

Larson,R., and Messing,J. (1982) *Nucleic Acids Res.,* **10**, 39-49.

Lee,K.H., Jones,R.A., Dalby,A., and Tsai,C.Y. (1976) *Biochem. Genet.,* **14**, 641-650.

Lerner,M.R., Boyle,J.A., Mount,S.M., Wolin,S.L., and Steitz,J.A. (1980) *Nature,* **283**, 220-224.

Lewis,E.D., Hagen,G., Mullins,J.I., Mascia,P.N., Park,W.D., Benton,D., and Rubenstein,I. (1981) *Gene,* **14**, 205-215.

Maniatis,T., Hardison,R.C., Lacy,E., Lauer,J. O'Connell,C., Quon,D., Sim,G.K., and Efstratiadis,A. (1978) *Cell,* **15**, 687-701.

Maxam,A.M., and Gilbert,W. (1977) *Proc. Natl. Acad. Sci. USA,* **74**, 560-564.

Messing,J., Crea,R., and Seeburg,H. (1981) *Nucleic Acids Res.,* **9**, 309-321.

Messing,J., and Vieira,J. (1982) *Gene,* **19**, 269-276.

Park,W.D., Lewis,E.D., and Rubenstein,I. (1980) *Plant Physiol.,* **65**, 98-106.

Proudfoot,N. (1980) *Nature,* **286**, 840-841.

Righetti,P.G., Gianazza,E., Viotti,A., and Soave,C. (1977) *Planta,* **136**, 115-123.

Rogers,J., and Wall,R. (1980) *Proc. Natl. Acad. Sci. USA,* **77**, 1877-1879.

Rubenstein,I. (1982) in Sheridan,W. (ed.), *Maize for Biological Research: A Special Publication of the Plant Molecular Biology Association,* University Press, University of North Dakota, Grand Forks, ND, pp. 189-195.

Sanger,F., Nicklen,S., and Coulson,A.R. (1977) *Proc. Natl. Acad. Sci. USA,* **74**, 5463-5467.

Stewart,J.W., Sherman,F., Shipman,N.A., and Jackson,M. (1971) *J. Biol. Chem.,* **246**, 7429-7445.

Sun,S.M., Slightom,J.L., and Hall,T.C. (1981) *Nature,* **289**, 37-41.

Vieira,J., and Messing,J. (1982) *Gene,* **19**, 259-268.

Wall,J.S. (1964) in Shutz,H.W., and Anglemier,A.F. (ed.), *Proteins and Their Reactions, Symposium on Foods,* Avi Publishing Co., Westport, pp. 315-341.

Wells,R.D., Goodman,T.C., Hillen,W., Horn,G.T., Klein,R.D., Larson, J.E., Muller,U.R., Nevendorf,S.K., Panoyotatos,N., and Stirdivant,S.M. (1980) *Prog. Nucleic Acids Res. Mol. Biol.,* **24**, 167-267.

Wienand,U., Langridge,P., and Feix,G. (1981) *Mol. Gen. Genet.,* **182**, 440-444.

Note added in proof

While this paper was in preparation, Pedersen *et al.* (*Cell,* 1982, **29**, 1015-1026) reported a complete zein genomic sequence, λZG99, and a partial zein cDNA sequence, pZ19.1. No intron is found in their genomic sequence, either. Although their genomic sequence does not have the 96-bp sequence in duplication, their cDNA sequence contains a similar tandem duplication to the one we showed here in Z4, with a single base difference at codon 127 (AGG in Z4 and AGA in λZG99). The flanking regions of the two genomic clones are more homologous to each other than their coding sequences. Regarding the tandem duplication, Heidecker and Messing (manuscript in preparation) have obtained a cDNA sequence that is 99% homologous to Z4 and contains the identical duplication.