The EMBO Journal Vol.1 No.12 pp.1589-1594, 1982

A homologous repetitive block structure underlies the heterogeneity of heavy and light chain zein genes

A.Spena, A.Viotti[1], and V.Pirrotta*

European Molecular Biology Laboratory, Postfach 102209, D-6900 Heidelberg, FRG, and [1]Istituto Biosintesi Vegetali, CNR, Via Bassini 15, 20133 Milano, Italy

Communicated by V.Pirrotta
Received on 19 October 1982

Heavy and light chain zein genes from maize, affected by different regulatory loci, are related in sequence and structure. The two kinds of genes code for a signal peptide, a head region of 67 amino acids, six and a half to eight repetitive blocks of ~20 amino acids each and a short tail piece. This block structure would allow inter or intragenic recombination giving rise to heterogeneous zein genes. Length variation in zein polypeptides is also due to the occurrence of termination mutations within some genes. Homology between heavy and light chain zein genes extends to the flanking sequences where a short region at the 5' end of the transcript can base pair with the 3' end and may have regulatory implications.

Key words: gene family/genomic sequence/periodic structure/inverted repeats

Introduction

Zeins are a family of proteins which accumulate in the endosperm of maize seeds where they serve as the principal form of nitrogen storage. Although they are very rich in glutamine and other polar amino acids, they contain very few charged residues and are overall highly hydrophobic and characteristically alcohol soluble. The zein polypeptides are similar in sequence and mol. wt. but they can be resolved by size into two classes of 19 000 and 22 500 apparent mol. wt. (Burr and Burr, 1976) and by isoelectric focusing into at least 20 distinct components (Righetti *et al.*, 1977).

The zeins are encoded by a family of ~120 genes of related sequence (Viotti *et al.*, 1979; Hagen and Rubenstein, 1981). Experiments with zein cDNA clones show that, in general, the high and low mol. wt. zein genes are distinguishable and non-cross-hybridising under mildly stringent conditions. There are at least two subclasses of heavy chain zein genes, coding for mature zeins of 22 000-23 000 apparent mol. wt. and two or probably three classes of light chain genes for zeins of 19 000-21 000 apparent mol. wt. (Viotti *et al.*, 1982; Burr *et al.*, 1982). The heavy and light chain genes are distinct in their chromosomal location: the light chains are on chromosomes 4, 7, and 10 while at least some heavy chains are at a different locus on chromosome 4 (Viotti *et al.*, 1982). Both kinds of zein genes are developmentally regulated. Their expression becomes detectable ~14 days after pollination and continues for ~40 days. A number of mutations affect this developmental program. Among these, the *opaque*-2 mutation inhibits preferentially the transcription of the heavy chain zein genes while the *opaque*-7 mutation affects the light chain zeins (Di Fonzo *et al.*, 1977; Burr and Burr, 1982). However, this inhibition is generally not complete and varies with the genetic background of the maize line studied.

*To whom reprint requests should be sent.

During the development of the seed, zeins are synthesized on polysomes attached to a specialised part of the endoplasmic reticulum. Secretion of the nascent protein across the reticular membrane is accompanied by the proteolytic cleavage of a signal peptide sequence at the amino terminus of the native protein which lowers the mol. wt. by ~2000 daltons (Larkins *et al.*, 1979).

The mature zeins accumulate in vesicles formed by the endoplasmic reticulum and, being insoluble, combine to form large aggregates constituting the so-called protein bodies (Burr and Burr, 1976). It is not known whether the zeins in the protein bodies assemble in specific structures. It is clear, however, that deficiencies of either heavy or light chain zeins, such as result from mutations at the *opaque*-2 or *opaque*-7 loci, affect the appearance of the seed and probably also its resistance to humidity and fungal infections (Lambert *et al.*, 1969) and might therefore have a structural basis.

Results

zA1, a heavy chain zein gene

A library of maize genomic clones was constructed by ligating a partial *Bam*HI digest of maize DNA to the λL47 vector of Loenen and Brammar (1980). We screened 200 000 plaques using cloned heavy chain or light chain zein cDNAs (Viotti *et al.*, 1982) and obtained both kinds of genomic clones. In this paper we describe clone zA1, containing a heavy chain zein gene and present conclusions obtained by comparing its structure with that of light chain zeins. A clone containing a cluster of light chain zein genes will be described elsewhere (Spena, Viotti, and Pirrotta, in preparation).

Clone pcM1, used as a probe for heavy chain sequences,

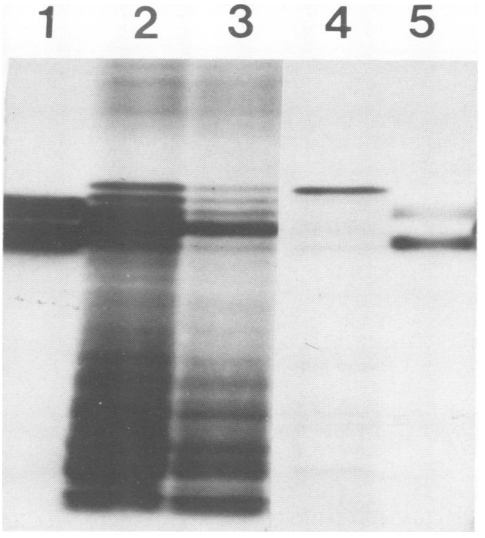

Fig. 1. *In vitro* translation of pcM1 selected mRNAs. Total endosperm RNA or hybrid selected RNA were translated in a wheat germ cell-free system in the presence of [^{3}H]leucine as described by Viotti *et al.* (1982). The products were analysed by SDS-PAGE and visualised by fluorography. **Lanes 1** and **5**: mature zein labelled by ^{3}H-dansylation. **Lanes 2** and **3**: translation products of total endosperm RNA. **Lane 4**: translation products of RNA selected by hybridisation to pcM1 DNA.

contains an insert of 1059 bp and is therefore a full length or nearly full length cDNA. It is classified as a heavy chain sequence because it hybridises to total zein mRNA from wild-type but not from *opaque*-2 mutants and selects mRNAs which translate *in vitro* into heavy chain polypeptides (Burr and Burr, 1982 and Figure 1). Restriction mapping, Southern blot analysis, and heteroduplex studies (data not shown) indicate that the genomic clone zA1 contains a region homologous and co-linear with pcM1 over at least 900 bp, with no evidence of intervening sequences. Clone zA1 does not hybridise to our light chain cDNA clones and has only one region of hybridisation to endosperm poly(A)+ RNA, corresponding to the region homologous to pcM1. A restriction map of the two clones as well as the strategy adopted to determine their nucleotide sequence are shown in Figure 2.

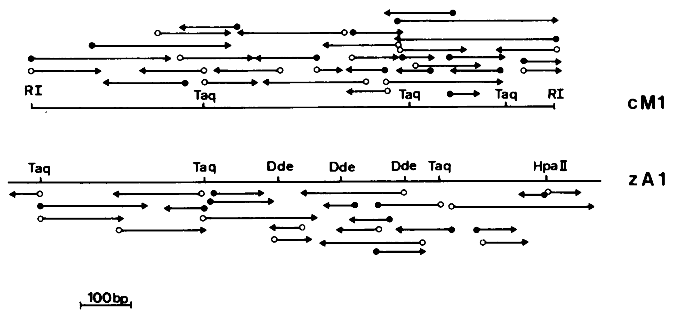

Fig. 2. Restriction map of the zA1 gene region and of the pcM1 clone. The strategy employed for sequencing by the Maxam-Gilbert method is shown below each map. Solid circles indicate 5′-labelled ends, open circles 3′-labelled ends. Nearly all sequences were determined on both strands and most, particularly the region containing the nonsense mutation in zA1, were sequenced several times using the G, G+A, A>C, T, T+C, and T reactions.

Comparison of the heavy chain cDNA and genomic sequences

The cDNA clone contains a long open reading frame beginning with an ATG near one end and ending with a TAG after 266 codons (Figure 3). A partial amino acid sequence determined on total heavy chain zein begins with S/FIIPQSSLAPSA... (Handa *et al.*, 1979) which corresponds to our sequence beginning with codon 22 after the initial ATG. The first 21 amino acids in fact constitute a signal peptide which is known to be cleaved off to give the mature zein. The sequence of pcM1 predicts a mol. wt. of 28 954 for the total polypeptide or 26 829 for the mature protein.

These mol. wts. are considerably higher than those previously reported (19 000 and 22 500 for light and heavy chain zeins, respectively, Burr and Burr, 1976) but agree with

```
AATACCATATAATCCGATAAAGTTGATAACTAAATCGAGTAGATGCCATATCATCTATATCTTATCTATTGTTTGGAAAAAATACAAAATCCAAAAAAAAATATTTGAGACCTCACCTATA
    .     -200      .     .     .     .     -150     .     .     .     .     -100

                      start        cM1                                                          M  A  T  K  V  L  A  L
TAAATAGCTCCCATATCAGTAGTTAATCCATCACCCATAATATTTTGAGCATTCAGAAACACACCAAGCGAAGCTCACTAGCAACGACCTAACAACAATGGCTACCAAGGTATTAGCCCT
    .     .     .     .     -50     .     .     AC.     .T     1     . A

 L  A  L  L  A  P  F  V  S  A  T  N  A  F  I  I  P  Q  C  S  L  A  P  S  A  I  I  P  Q  F  L  P  P  V  T  S  M  G  F  E
CCTTGCGCTTCTTGCCCCTTTTGTGAGCGCAACAAATGCGTTCATTATTCCACAATGCTCACTTGCTCCTAGTGCCATTATTCCACAGTTCCTCCCACCAGTTACTTCAATGGGCTTCGA
    .     .T     50     .     .     .     .     100  A  .     .     .     .

 H  L  A  V  Q  A  Y  M  L  Q  Q  A  L  A  A  S  V  L  Q  Q  P  I  D  Q  L  Q  Q  Q  S  L  A  H  L  T  I  Q  T  I  A  T
ACACCTAGCTGTGCAAGCCTACATGCTACAACAAGCGCTTGCAGCGAGCGTCTTACAACAACCAATTGACCAATTGCAACAACAATCCTTGGCACATCTAACCATACAAACCATCGCAAC
   150     .     G .     .     G A.     200     .A     .     .     .     250     .

 Q  Q  Q  Q  Q  F  L  P  A  L  S  Q  L  A  V  V  N  P  V  A  Y  L  Q  Q  Q  L  L  A  S  N  P  L  A  L  A  N  V  A  A  Y
GCAACAGCAACAACAGTTCCTACCAGCACTGAGCCAACTAGCTGTGGTGAACCCTGTCGCCTACTTGCAACAGCAGCTGCTTGCATCCAACCCACTTGCTCTGGCAAACGTAGCTGCATA
    .     .     .     300  A  .     .     .     .     350     .     .     .

 Q  Q  Q  Q  L  Q  Q  F  L  P  A  L  S  Q  L  A  M  V  N  P  A  A  Y  L  Q  Q  Q  Q  L  L  S  S  S  P  L  A  V  V  N  A
CCAACAACAACAATTGCAACAGTTTCTACCAGCGCTCAGTCAACTAGCCATGGTGAACCCTGCCGCCTACCTACAACAGCAACAACTGCTTTCATCTAGCCCTCTCGCTGTGGTTAATGC
     CAA     . G     .G     .     .     .     450     .     .     .     .     G  500

 P  T  Y  L  Q  Q  Q  M  L  E  Q  I  V  P  A  L  T  Q  L  A  V  A  N  P  A  A  Y  L  Q  Q  L  L  P  F  N  Q  L  T  V  S
ACCTACATACCTGCAACAACAGATGCTGGAACAGATTGTACCAGCTCTGACTCAGCTAGCTGTGGCAAACCCTGCTGCCTACTTGCAACAGCTGCTTCCATTCAACCAACTGACTGTGTC
    .     .     AT   . C     .     550     .     .     .     .     600     .     .

 N  S  A  A  Y  L  Q  Q  R  Q  P  L  L  N  P  L  A  V  A  N  P  L  V  A  A  F  L  *  Q  Q  Q  L  L  P  Y  N  Q  F  S  L
GAACTCTGCTGCGTACCTACAACAGCGACAACCATTACTTAATCCACTGGCAGTGGCTAATCCATTGGTCGCTGCCTTCCTATAGCAGCAACAATTGCTGCCATACAACCAGTTCTCTTT
    .     .     650  AG  .     . A A   C.A  C     .     G700  C  .     . A    .G     .

 M  N  P  A  L  S  W  Q  Q  P  I  V  G  G  A  I  F  *
GATGAACCCTGCCTTGTCGTGGCAGCAACCCATCGTTGGAGGTGCCATCTTTTAGATTACATATGAAATGTACTTGATAATGGTGCCCTCATACCGGCATGTGTTTCCTAGAAATAATCA
   750     .     .     .     .     800     G     C .     .     .     850     .A

              end
ATATATTGATTGAGATTTATCTGGATATATTTTTGAACTATGTCTACTATATAAATAATTGAAAGCATCATATCGTAATTATAAACTCATGGTTGGTTAATACATGATAATACAATATTA
    .     .     .     900     .     .     .     .     950     .     .     .

ACTCACATCCAATGTGAG
    .     1000
```

Fig. 3. Nucleotide sequence of gene zA1. The nucleotides are numbered beginning with the A of the ATG initiation codon. Nucleotide differences and the additional codon at position 390 in pcM1 are indicated below the zA1 sequence. Terminator codon, TATA box, and polyadenylation signal are boxed. Sequences at the 5′ and 3′ ends of the gene which are involved in the inverted terminal repetition are underlined.

more recent determinations which show that the earlier values were based on calibrations with anomalously migrating proteins (Vitale and Viotti, in preparation). The new determinations give values of 23 000 – 27 000 which agree also with those predicted by the sequence of light chain zein genes (Geraghty *et al.*, 1981; Pedersen *et al.*, 1982).

The pcM1 sequence is essentially identical to that of the zA1 gene. Less than 4% of the nucleotides are variant, in many cases without affecting the amino acid sequence. The two sequences have two important differences. One is an insertion of an extra codon in pcM1 which has five CAAs while zA1 has four around codon number 130. The other change in zA1 turns a CAG into a TAG, introducing a terminator codon at position 236 in the gene. The zA1 mRNA would therefore code for a zein polypeptide of only 23 210 daltons, making it equivalent in size to a light chain zein. Such terminator mutations resulting in truncated zeins are apparently fairly frequent. We have found another such prematurely terminated gene among light chain sequences determined so far. Both heavy and light chain zeins are rich in codons such as CAA or CAG which can be converted into terminators. Furthermore, since zein genes belong to a multigene family and the function of their product is to store nitrogen, we suppose that they are not individually subject to strong selection and that a shorter protein can still contribute storage potential. Prematurely terminated genes would be expected to accumulate as they do in the leghemoglobin multigene family of soy bean (Brisson and Verma, 1982).

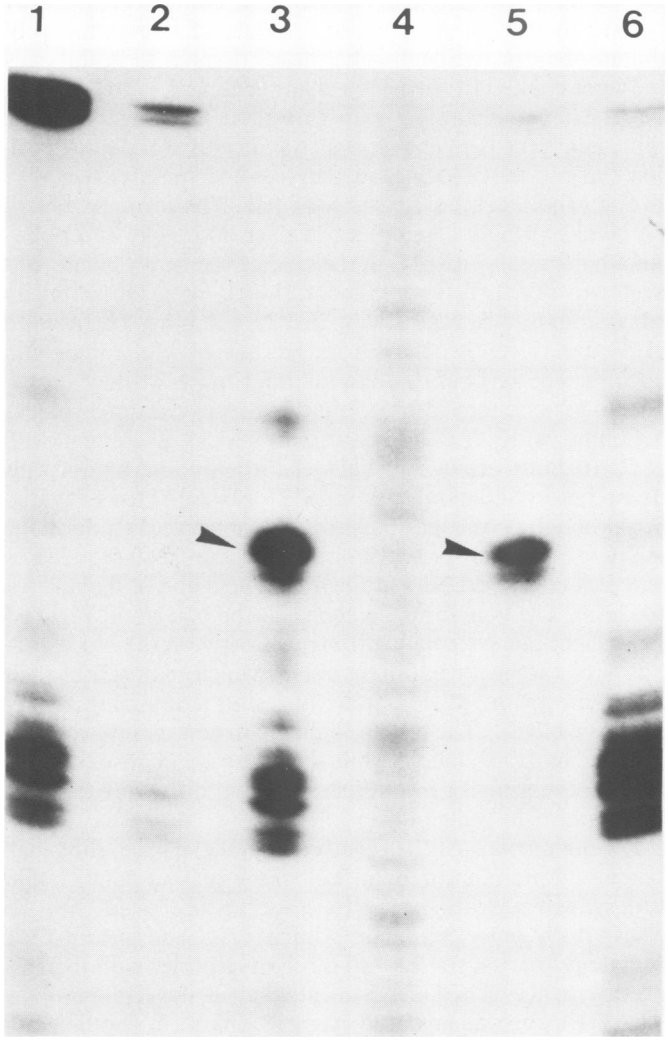

Fig. 4. S1 mapping of the zA1 gene. The *Taq-Taq* 321-bp fragment containing the 5′ end of the zA1 gene was labelled with polynucleotide kinase and hybridised with endosperm poly(A)$^+$ RNA. **Lanes 1, 3,** and **6** were mildly S1 treated and represent: **lane 1,** non-hybridised *Taq* fragment, **3,** hybridised with wild-type RNA, and **6,** hybridised with *opaque*-2 RNA. **Lanes 2** and **5** were more stringently S1 treated: **2,** hybridised with *opaque*-2 RNA and **5,** hybridised with wild-type RNA. **Lane 4** contains an end-labelled DNA fragment of known sequence partially degraded by the Maxam-Gilbert A + G reaction as a size marker. The arrowheads indicate the fragments protected against stringent S1 digestion.

That these are probably not pseudogenes in the sense of being inactive and non-functional is suggested by the observation of minority zein polypeptides of different mol. wt. Figure 1 shows, for example, the product of *in vitro* translation of RNA selected by pcM1. Below the main product of heavy chain zein can be seen several fainter bands of lower mol. wt. all of which are soluble in ethanol. One of these, migrating as a light chain zein, might correspond to the product of the zA1 gene. Although these bands were observed in different *in vitro* translation systems and with different mRNA preparations, we cannot rule out the possibility that they are due to translation artefacts.

Downstream of its premature termination codon, the zA1 gene continues in almost perfect correspondence with the pcM1 sequence. The normal termination codon is a TAG at position 265 (266 in pcM1) and is followed by 83 nucleotides before the onset of the poly(A) tail in the cDNA clone. The polyadenylation site is preceded by the usual eukaryotic polyadenylation signal sequence AATAAA (Proudfoot and Brownlee, 1974) whose first A is located 27 nucleotides from the poly(A) tail. In the zA1 clone and in several light chain genes, this sequence is slightly different: AAATAA.

The 5′ ends of the two clones are homologous up to nucleotide −59 where they diverge due to a cloning artefact in pcM1. To determine the position of the true 5′ end of the mRNA, we used the S1 mapping procedure (Berk and Sharp, 1977). A *Taq* fragment which starts at nucleotide 140 and includes the 5′ end of the gene was end labelled and hybridised to zein mRNA from wild-type or from an *opaque*-2 mutant. Figure 4 shows that no detectable protection against S1 digestion is obtained with the *opaque*-2 mRNA, indicating that all sequences homologous to zA1 are controlled by *opaque*-2. After S1 treatment, wild-type mRNA protects a fragment of 215 nucleotides with an uncertainty of four nucleotides. This places the 5′ end of the mRNA, and presumably the start of transcription, approximately at position −72. Upstream of this site, a TATAAAT sequence whose first T is 28 nucleotides away from the presumed transcription start, constitutes a classical Goldberg-Hogness box (Gannon *et al.*, 1979).

The 5′- and 3′-flanking sequences of the zA1 gene have another interesting property. When folded back, the two ends of the gene have a considerable degree of inverted homology allowing the base pairing of 20 of 25 bases at the 5′ side (positions −34 to −59) with 20 of 28 bases at the 3′ side (positions 843 – 870). A similar but much shorter inverted homology (7 – 9 bases) is also found at the two ends of all light chain zein genes published so far. The significance of this inverted homology is not clear but that the two ends of the mRNA can base pair is shown by the production of artefacts at the 5′ end of cDNA molecules as will be described elsewhere (Spena, Viotti, and Pirrotta, in preparation). Base pairing between the 5′ and 3′ ends of the mRNA would surely affect its translational potential and may be involved in translational regulatory mechanisms.

Comparison of heavy and light chain zein genes

The signal peptide. The coding region of the zA1 gene has interesting features which are best understood by comparing it with the coding region of a light chain zein gene. Several very similar light chain genes are now available in the literature and from our own work (Geraghty *et al.*, 1981; Pedersen *et al.*, 1982). In the present discussion we will use, for comparison, the sequence from a genomic clone containing a cluster of light chain zeins (Spena, Viotti, and Pirrotta, in preparation). Figure 5 compares the structure of the A1 gene, translated into the one letter amino acid code, with that of the light chain gene. Both genes begin with signal peptides of 21 amino acids very similar in structure and sequence. Both follow closely the general rules governing signal peptides (Inouye and Halegoua, 1979): they are 18–24 amino acids long, contain a basic residue near the NH_2 end (lysine at position 4), followed by a highly hydrophobic stretch at the end of which is frequently found a proline or glycine. The cleavage occurs after an amino acid with short side chain (Ala at position 21).

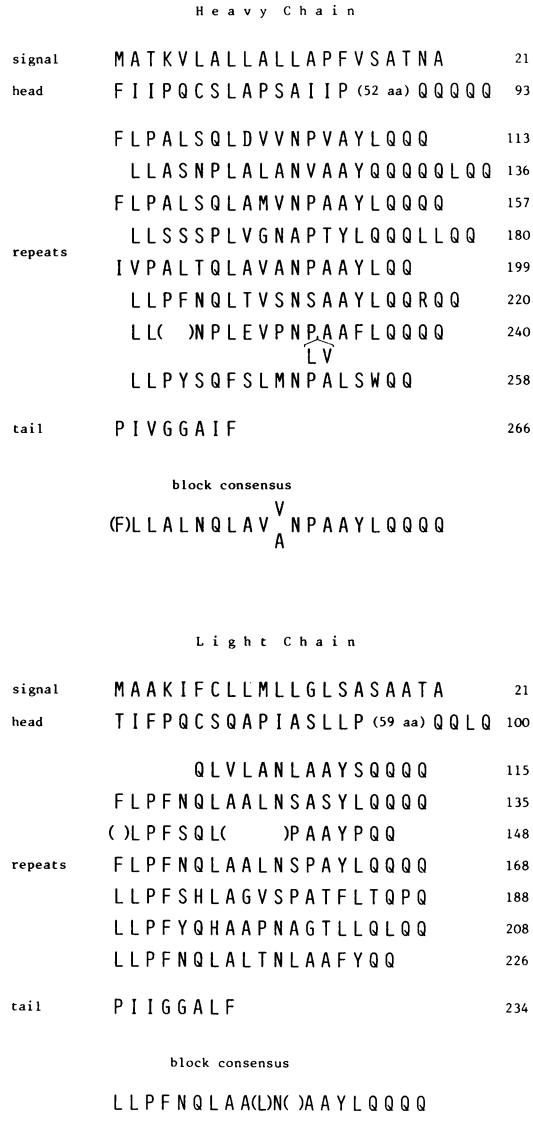

Fig. 5. Block structure of heavy and light chain zein polypeptides. The amino acids are represented by the one letter code, numbered from the initiation codon. Gaps, indicated by parentheses, and an insertion of two amino acids are arranged to emphasise the repetitive block homology whose concensus sequence is shown below.

The 'head' domain. Sequence homology between the heavy and light chain genes continues past the proteolytic cleavage site and the amino terminus of the mature protein. The similarity becomes harder to detect after some 40 amino acid residues except for certain repeated motifs such as the run of glutamines at position 74–76. This region, until amino acid 88 (nucleotide 264), constitutes the 'head' domain and contains all or almost all the charged residues found in the mature polypeptide as well as one cysteine (two in light chain zeins).

The repetitive blocks. The remaining two thirds of the gene both in the heavy and light chain zeins has a periodic structure constructed of blocks of ~20 amino acids which repeat a consensus sequence common to both kinds of zeins. Similar blocks have been reported for light chain zeins by Geraghty *et al.* (1981) and by Pedersen *et al.* (1982) who have chosen different cyclic permutations of the block sequence as their repeat unit. Traces of the repeat are also discernible in the 'head' domain suggesting that this too was derived from the basic building block. Transitions between blocks are marked by runs of glutamines. As Figure 5 shows, a run of glutamines also terminates the 'head' domain in both heavy and light chains. In the heavy chain gene, this is followed by the first block beginning with FLP... at position 94. In the light chain gene the periodic structure begins at the same position but the first block is incomplete and initiates half way through the block consensus sequence. The arrangement of the blocks in Figure 5 emphasises their homology and points out a curious alternation, in the case of the heavy chain, between two slightly different types of block, one, containing an extra amino acid, begins with FLPA, the other begins with LLA and is followed by a longer tail of glutamines and leucines. The germ of this alternation, though without the extra amino acid, may also be found in the light chain as blocks beginning with FLP alternate with blocks beginning with LLP. In both heavy and light chains, the alternation fails after a few cycles and the last three blocks are of the LL variety. Although occasional blocks are abnormal, with deletions or insertions of one or two amino acids, the block sequence is very well conserved both at the amino acid and at the nucleotide level. The consensus sequences at the bottom of Figure 5 show slight but systematic differences between the block consensus of the heavy and light chains which may reflect their evolution from a common ancestor containing fewer repetitive blocks. The present light chain gene contains about six and a half such blocks while the heavy chain gene contains eight, accounting for most of the mol. wt. difference between heavy and light zeins.

The tail piece. At the carboxyl end of the gene, the block structure is terminated by a 'tail' piece of non-polar amino acids which are also strongly conserved between the two genes. Both genes terminate with a TAG codon and their homology persists for 20–30 more nucleotides with increasing divergence. They are similar again in the region of the AAATAA polyadenylation signal which in both cases precedes the poly(A) addition site by 27 nucleotides.

Homology in the promoter region

Of considerable interest is the homology at the 5′ end of the genes, preceding the translated region. Figure 6 shows the nucleotide sequences of zA1 and of a light chain gene zE19, aligned to emphasise the homology in this region. The zE19 gene, whose complete sequence will be reported elsewhere, is >95% homologous to the light chain gene of Pedersen *et al.*

(1982). Upstream of the initiation codon ATG, the A1 gene has a sequence of 99 nucleotides to the TATAAATA box. The light chain gene has a largely homologous sequence if allowance is made for two small deletions, one of which eliminates the reduplication of a AAGC present in the A1 gene. S1 mapping places the 5′ end of the A1 mRNA around position −72. Since this is at about the canonical distance from the TATAAATA box (26−32 bp) we suppose that it corresponds to the site of transcription initiation. In the light chain gene, Pedersen *et al.* (1982) place the 5′ end of the mRNA a few nucleotides to the right, at a distance of 34 nucleotides from the TATAAATA box.

Homology continues upstream of the TATAAATA box into a region characterised by a series of stutter-like repetitions. Differences between the two genes in this region are frequently due to the presence or absence of a repetition.

The region of recognizable homology between the two genes extends to ~120 nucleotides upstream of the transcription initiation site. Further upstream, although different light chain zeins continue to be almost identical with each other for >400 nucleotides, the homology with the heavy chain gene ceases. The 120-bp region preceding the transcription start is generally thought to contain the essential features of eukaryotic promoters. Among these are the TATA box and the so-called CAT box, a GCAAT sequence situated 70−80 nucleotides upstream of the transcription start (Benoist *et al.*, 1980). Neither heavy nor light chain zeins have a GCAAT in this region but both have a series of two or three G/CAAAATs ~70 bp from the transcription start, which might be its functional equivalent. Another similar sequence (CCAT) is found 99 bp from the start site.

Both heavy and light chain genes are developmentally regulated, presumably through specific regulatory sequences. The two kinds of gene are however controlled by different loci: *opaque*-2 for heavy chain genes and *opaque*-7 for light chain genes. It is possible that the sequence divergence upstream of the promoter region reflects the two different regulatory systems.

Discussion

Our results show that heavy and light chain zein genes are very similar both in structure and in sequence. Both kinds of zeins are constructed of a signal peptide, a head domain, a repetitive domain composed of a variable number of block units, and a short tail piece.

The repetitive blocks which make up the bulk of the protein clearly imply repetitions in the three-dimensional structure of the protein and are most likely folded in an interacting zig-zag arrangement placing the glutamine runs at the hinges between blocks. However, given the scarcity of structural information on hydrophobic, non-globular proteins, structural predictions are at present not very secure. We suppose, however, that the head regions, which contain several charged and polar residues (including one or two cysteines) might congregate together. The formation of zein dimers which dissociate in the presence of mercaptoethanol has been observed by Vitale *et al.* (1982).

Although the head region is very divergent from the block consensus, traces of the repetitive block sequence may be detected in it and it is likely to have originally derived by a specialization of two or three such blocks in a common ancestor of the present zeins. This ancestor must itself have originated by internal tandem duplication within a gene containing only one repeat unit. The divergence between heavy and light chain genes, which involves both the number of blocks and small but consistent differences in the block consensus sequence, is most easily explained if it occurred before the full number of repeats had been reached and before the colonisation of the several chromosomal sites.

At present, at least four different clusters of zein genes on three different chromosomes can be identified by *in situ* hybridisation. There are at least five subfamilies of zein genes as defined by cross-hybridisation: three light chain and two heavy chain families (Viotti *et al.*, 1982). These however do not strictly correspond to the chromosomal loci.

It is not clear whether the heavy and light chain zeins have specific roles in the structural organization of protein bodies. In addition to the mol. wt. difference, due to the presence of an additional one and a half block units in the heavy chain, the two kinds of zeins have a number of systematic differences in their amino acid sequence. These sequence differences and the consequent structural differences might be more important than the presence or absence of an extra block. At the present time, however, there is no evidence for light chain or heavy chain specific interactions.

The repetitive block structure of the zein genes and their occurrence in clusters suggests the possibility of unequal crossing over between genes and the generation of variant zeins with a broad range of mol. wts. A gene apparently due to recombination of this sort has been detected already. Pedersen *et al.* (1982) have reported a zein clone containing an almost exact internal repetition of 96 nucleotides involving one and a half repetitive blocks after which the sequence returns to continuous alignment with the typical light chain sequence. An additional contribution to the generation of mol. wt. variants is the frequent occurrence of nonsense mutations resulting in premature termination of translation such as occurs in the zA1 gene. A more significant classification of the zein genes might be based not on size of the polypeptide encoded but on sequence homology of the genes

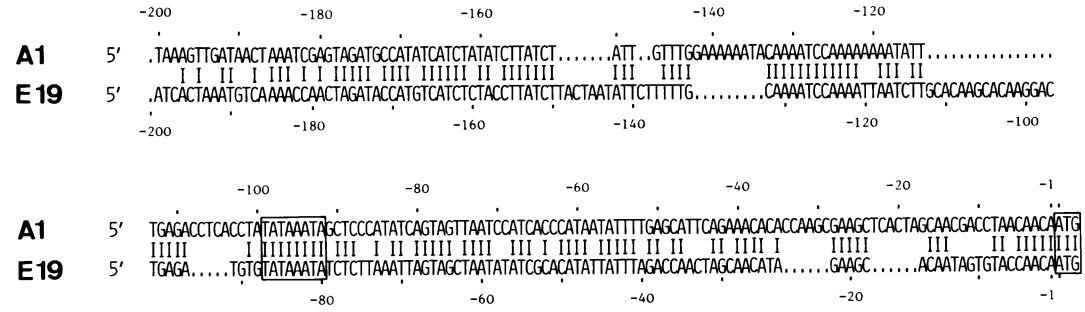

Fig. 6. Comparison of the 5′-flanking sequences of heavy and light chain zein genes. The sequences are numbered starting with nucleotide −1 preceding the ATG translation initiation codon. Dots indicate gaps introduced to emphasise homology.

and flanking sequences, on *in situ* localisation, and on common control.

It remains to be seen whether the differences between the classes of zeins is a functional one and why separate mechanisms have evolved to regulate their expression since they are expressed in the same place and at the same time. Perhaps some zein genes function also in other tissues or at other developmental stages.

Materials and methods

Construction of maize genomic clone library

Maize DNA was extracted and purified from unfertilized ears of the W64 A line as described by Viotti *et al.* (1982). DNA fragments produced by partial *Bam*HI digestion were sized on a NaCl gradient (Lin *et al.*, 1980) and cloned in the λL47 vector of Loenen and Brammar (1980). After *in vitro* packaging as described by Scherer *et al.* (1981), the phages were plated on a bacterial host lysogenic for phage P2, which allows the selection of recombinant phages. About 200 000 plaques were screened by the method of Benton and Davis (1977), using cloned cDNA sequences (Viotti *et al.*, 1982). For detailed analysis and sequencing, DNA fragments of interest were subcloned into plasmid vectors.

DNA sequencing

Restriction fragments, purified by gel electrophoresis or by subcloning, were labelled at the 5′ or 3′ ends with polynucleotide kinase or DNA polymerase (Klenow), respectively. Fragments labelled at a single end were obtained by strand separation (Szalay *et al.*, 1977) or by a second restriction cut. The chemical degradation reactions were performed according to Maxam and Gilbert (1980) and Rubin and Schmid (1980).

S1 mapping

We followed the procedure of Berk and Sharp (1977). About 5 μg of the *Taq-Taq* 321-bp fragment containing the 5′ end of the zA1 gene were labelled with polynucleotide kinase, and hybridized with 20 μg of endosperm poly(A)$^+$ RNA in 60% formamide, 0.4 M NaCl, 10 mM EDTA, and 0.1 M PIPES (piperazine-N,N′-bis 2-ethane sulfonic acid) pH 6.8 for 14 h at 52°C. S1 digestion was carried out at 37°C in buffer containing 30 mM Na acetate, pH 4.5, 0.3 M NaCl, 3 mM $ZnCl_2$, and 60 μg sonicated, denatured salmon sperm DNA. Two degrees of stringency were used: either with 10 units of S1 for 1 h or with 20 units for 2 h. The products were then analysed on a sequencing gel containing 6% acrylamide and 8 M urea.

Acknowledgements

We are grateful to R.Lacatena and L.Bougueleret for help in the sequence analysis. A.S. is indebted to G.Cesareni for room, board, and helpful criticism. A.V. was the recipient of an EMBO short term fellowship and A.S. was supported by a grant from the Commission of the European Communities.

References

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

Benton,W.D., and Davis,R.W. (1977) *Science (Wash.)*, **196**, 180-182.

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

Brisson,N., and Verma,D.P.S. (1982) *Proc. Natl. Acad. Sci. USA*, **79**, 4055-4059.

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

Burr,B., and Burr,F.A. (1982) *J. Cell Biol.*, **94**, 201-206.

Burr,B., Burr,F.A., St.John,T.P., Thomas,M., and Davis,R.W. (1982) *J. Mol. Biol.*, **1**, 33-49.

Di Fonzo,N., Fornasari,E., Salamini,F., and Soave,C. (1977) *Maydica*, **22**, 77-78.

Gannon,F., O'Hare,K., Perrin,F., LePennec,J.P., Benoist,C., Cochet,M., Breathnach,R., Royal,A., Garajoin,A., Cami,B., and Chambon,P. (1979) *Nature*, **278**, 428-434.

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

Hagen,G., and Rubenstein,I. (1981) *Gene*, **13**, 239-249.

Handa,A.K., Harmodson,M., Tsai,C.Y., and Larkins,B.A. (1979) *Plant Physiol.*, **63**, Suppl., 94.

Inouye,M., and Halegoua,S. (1979) *CRC Crit. Rev. Biochem.*, **7**, 339-371.

Lambert,R.J., Alexander,D.E., and Dudley,J.W. (1969) *Crop Sci.*, **9**, 242-243.

Larkins,B.A., Pedersen,K., Handa,A.K., Hurkman,W.J., and Smith,L.D. (1979) *Proc. Natl. Acad. Sci. USA*, **76**, 6448-6452.

Loenen,W., and Brammar,W.S. (1980) *Gene*, **20**, 249-259.

Lin,C.H., Tucker,P.W., Mushinski,J.F., and Blattner,F.R. (1980) *Science (Wash.)*, **209**, 1348-1353.

Maxam,A.M., and Gilbert,W. (1980) in Grossman,L., and Moldave,K. (eds.), *Methods in Enzymology*, Vol. **65**, Academic Press, NY, pp. 499-560.

Pedersen,K., Devereux,J., Wilson,D.R., Sheldon,E., and Larkins,B.A. (1982) *Cell*, **29**, 1015-1026.

Proudfoot,N.J., and Brownlee,G.G. (1974) *Nature*, **252**, 359-362.

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

Rubin,C.M., and Schmid,C.W. (1980) *Nucleic Acids Res.*, **8**, 4613-4619.

Scherer,G., Telford,J., Baldari,C., and Pirrotta,V. (1981) *Dev. Biol.*, **86**, 438-447.

Szalay,A.A., Grohmann,K., and Sinsheimer,R.L. (1977) *Nucleic Acids Res.*, **4**, 1569-1578.

Viotti,A., Sala,E., Marotta,R., Alberi,P., Balducci,C., and Soave,C. (1979) *Eur. J. Biochem.*, **102**, 211-222.

Viotti,A., Pogna,N., Balducci,C., and Durante,M. (1980) *Mol. Gen. Genet.*, **178**, 35-41.

Viotti,A., Abildsten,D., Pogna,N., Sala,E., and Pirrotta,V. (1982) *EMBO J.*, **1**, 53-58.

Vitale,A., Smaniotto,E., Longhi,R., and Galante,E. (1982) *J. Exp. Bot.*, **33**, 439-448.