The EMBO Journal Vol.1 No.11 1982 pp.1329–1335, 1982

# Sequence analysis and comparison of cDNAs of the zein multigene family

**Daniel E. Geraghty, Joachim Messing[1], and Irwin Rubenstein***

Department of Genetics and Cell Biology, and [1]Department of Biochemistry, University of Minnesota, St. Paul, MN 55108, USA

*Communicated by P. Starlinger*
*Received on 1 September 1982*

**The nucleotide sequence of two zein cDNAs in hybrid plasmids A20 and B49 have been determined. The insert in A20 is 921 bp long including a 5′ non-coding region of 60 nucleotides, preceded by what is believed to be an artifactual sequence of 41 nucleotides, and a 3′ non-coding region of 87 nucleotides. The B49 insert is 467 bp long and includes approximately one-half the protein coding sequence as well as a 3′ non-coding region of 97 nucleotides. These sequences have been compared with the previously published sequence of another zein clone, A30. A20 and A30, both encoding 19 000 mol. wt. zeins, have ~85% homology at the nucleotide level. The B49 sequence, corresponding to a 22 000 mol. wt. zein, has ~65% homology to either A20 or A30. All three zeins share common features including nearly identical amino acid compositions. In addition, the tandem repeats of 20 amino acids first seen in A30 are also present in A20 and B49.**
*Key words:* zein cDNAs/DNA sequence/multigene/repetitive structure

## Introduction

Multigene families are among the most intensively studied aspects of eucaryotic genomes. They have the potential to explain much about the structure and regulation of eucaryotic genes and have been invaluable in understanding the mechanisms of evolution. Much of this work has been done in animal systems while relatively few examples of multigene families have been well defined in plants. Of these few, storage proteins are among the best characterized due to their abundance and economic importance. The zein multigene family which represents the major storage proteins of the maize endosperm, has been under investigation in a number of laboratories.

Estimates of the number of genes in the zein family vary from a low of eight (Pederson *et al.*, 1980) to >100 (Viotti *et al.*, 1979; Wienand and Feix, 1980; Hagen and Rubenstein, 1981). When zein proteins are separated in SDS gels, two major classes of mol. wts. 19 000 and 22 000 are seen along with minor classes of mol. wts. 10 000 and 14 000 (Misra *et al.*, 1975; Soave *et al.*, 1976). Fifteen to twenty-five distinct protein spots can be seen on isoelectric focusing gels (Righetti *et al.*, 1977; Hagen and Rubenstein, 1980) and genetic data indicate at least 16 separate loci on maize chromosomes 4 and 7 (Soave *et al.*, 1981, 1982). Park *et al.* (1980) have divided the zein mRNAs into at least three subfamilies, each of which is defined by hybridization to one of three zein cDNA clones. We have recently published the DNA sequence of one of these clones, A30 (Geraghty *et al.*, 1981), and present here the sequence of the remaining two.

A20, the longer of these two clones, may represent a nearly full length copy of its corresponding mRNA. This sequence contains one long open reading frame coding for a protein of 240 amino acids. Upon comparison with protein sequence data, the first 21 amino acids appear to be the signal peptide shown to be present in most if not all zeins (Burr *et al.*, 1978; Larkins and Hurkman, 1978). In hybrid selection experiments the A20 clone binds mRNAs that are translated into the 19 000 mol. wt. class of zeins. The second clone, B49, contains a shorter cDNA insert which covers the 3′ non-coding regior. and approximately one-half the protein coding sequence. In hybrid selection experiments B49 binds mainly mRNAs coding for the 22 000 mol. wt. size class of zeins.

Where possible, the nucleotide sequences of these three cDNAs were compared. The 3′ non-coding regions of A20 and A30 are similar, while this region of B49 is much more divergent. While the amino acid sequence deduced from the nucleotide sequence of these three clones have diverged due to substitutions and insertions or deletions, the amino acid composition remains very similar. Both of the zein proteins coded for by the A20 and B49 clones show the repetitive amino acid sequence at their carboxy terminal portions first seen in A30.

We also include a correction to the previously published sequence of A30, deleting the ninth nucleotide from the 5′ end. This correction in the sequence extends the predicted length of the signal peptide as discussed below.

## Results and Discussion

The sequencing strategy for A20 and B49 is shown in Figure 1. Information from previous restriction mapping (Park *et al.*, 1980) was used to begin the sequencing. The sequence derived was then used to select other restriction sites to complete the sequencing. Nearly the entire sequence of

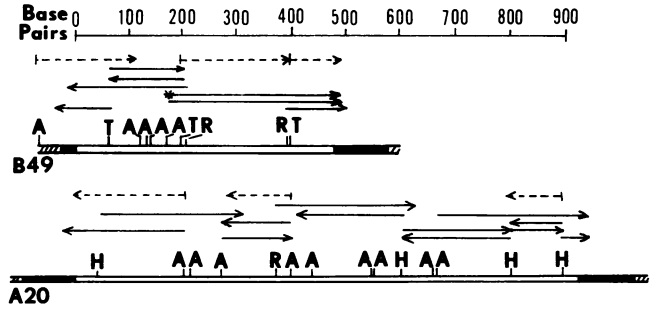


**Fig. 1.** Partial restriction map and sequencing strategy of A20 and B49. Clone A20 contains a 921-nucleotide insertion (open line) flanked by poly(A/T) tails (solid line) in the *Eco*RI site of plasmid PMB9 (hatched line). Clone B49 contains a 467-nucleotide insert and is drawn in the same manner as A20. Only the restriction sites used in end labeling or cloning and subsequent sequence analysis is shown. A = *Alu*I, H = *Hae*III, R = *Rsa*I, and T = *Taq*I. Solid arrows identify the direction and extent of sequences determined by the method of Maxam and Gilbert (1980). All fragments were 5′ end-labeled except for the B49 *Alu* fragment (asterisk) which was 3′ end-labeled. Dashed arrows identify subclones of A20 and B49 in M13mp7 sequenced by the dideoxy chain terminator technique.

*To whom reprint requests should be sent.

 0261-4189/82/0111-1329$2.00/0.

both strands of A20 and B49 was determined and the sequences at all restriction sites were overlapped. Also, all the restriction sites predicted from the nucleotide sequences have been confirmed.

The nucleotide sequence of the A20 cDNA clone is given in Figure 2 aligned with the nucleotide sequences of B49 and A30. The insert of A20 is 921 nucleotides in length and possibly covers the entire length of the mRNA including the 5′ non-coding region. The nucleotide sequence of the B49 clone is aligned to give maximum homology with A20 and A30 and consists of 467 nucleotides between the A/T tails. Considering the expected size of B49-like mRNA (~1000 nucleotides), this cDNA is about one-half the length of the mature message. The amino acid sequences predicted from the nucleotide sequences are shown aligned in Figure 3. The nucleotide and amino acid sequences of A20 and B49 have obvious homology with those of A30. They show the same general features of zein proteins including the amino acid composition and a signal peptide present in A20.

As mentioned in the Introduction, A20 hybrid selects the 19 000 mol. wt. class of zeins. However, the mol. wt. of the protein deduced from the A20 sequence, less the signal peptide, is 24 027 daltons. This same ambiguity was noted with the mol. wt. of the A30 coded protein and is seen in other clones sequenced in our laboratories. Evidently the apparent mol. wts. of zeins as usually determined by SDS-gel electrophoresis are underestimates by 4 – 5 kd. This is not surprising considering the hydrophobicity of zein proteins as compared to those proteins usually used as mol. wt. markers.

*Nucleotide and amino acid sequences*

When the nucleotide sequences of clones A20 and A30 were aligned and compared, 123 differences were observed: 120 base substitutions and three insertions and/or deletions (85% homology). Of the 106 substitutions in the coding region, 65 are transitions and 41 are transversions giving a ratio of 1.6 (0.5 expected on a random basis). The frequency of substitution is constant throughout the entire length of the A20 and A30 sequences although the changes tend to be somewhat clustered. In contrast, the types of changes are not evenly distributed; the transition/transversion ratio is 2.2 for the proximal two-thirds of the codings sequences, 1.0 for the terminal one-third, and 0.75 for the 3′ non-coding sequence. The coding region of clone B49 requires seven gaps to achieve the alignment shown in Figure 2, resulting in ~70% homology to either of the other sequences. The greater extent of divergence seen in the 3′ non-coding sequence leaves the overall homology at ~65%.

In the mature proteins, there are 48 amino acid substitutions between A20 and A30, the three insertions/deletions leaving A20 six amino acids longer. Even with this number of substitutions the overall amino acid composition changes very little (see Table I). Both zeins have the same number of non-polar residues and have very nearly equal amounts of leucine, glutamine, alanine, and proline which account for slightly more than 60% of the amino acid residues of these proteins. A similar comparison with B49 is difficult since this clone does not span the entire coding region. However, the amino acid composition of the portion presented is consistent with the others. This is seen most clearly with the four most abundant amino acids mentioned above. Together they comprise 60% of the B49 protein's residues and each individually comprises nearly the same percentage seen in A20 or A30. One unusual feature of the B49 protein is the presence of a tryptophan residue near the carboxy terminus. Zeins are low in tryptophan and lysine (Wall, 1964) and this is the only occurrence (excluding the signal peptide) of either amino acid in the zeins sequenced to date.

*Signal peptides*

Both the A20 and the corrected A30 sequences contain two potential ATG initiator codons, 21 and 12 codons preceding

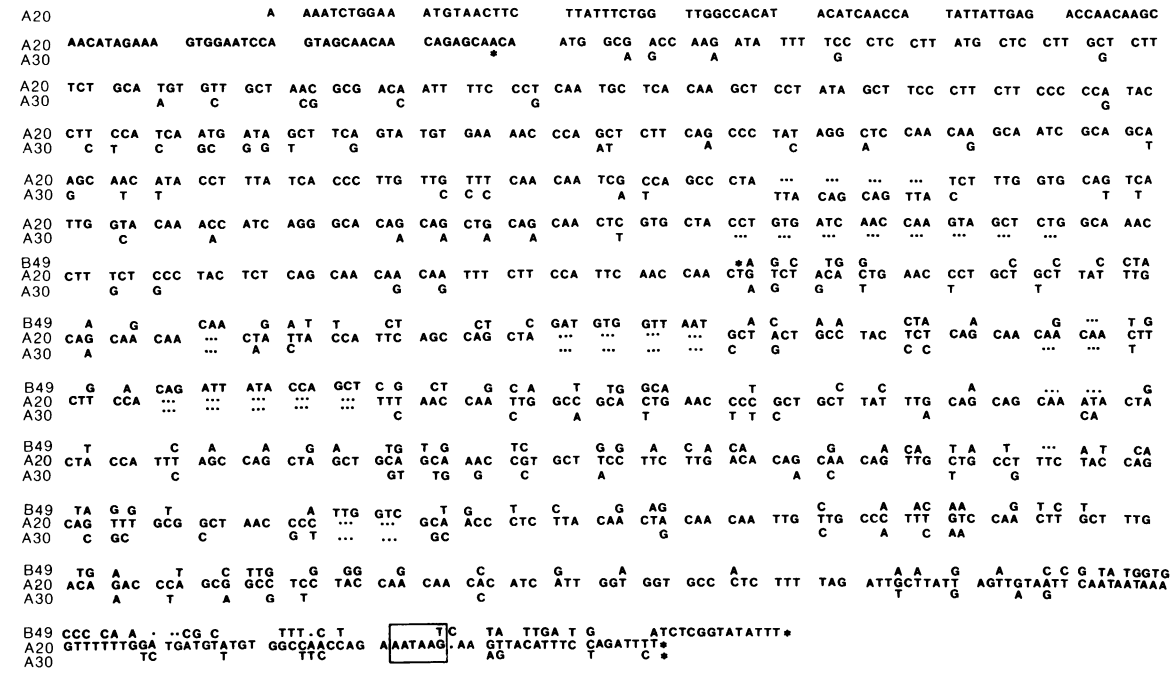

**Fig. 2.** The nucleotide sequences of the A20, A30, and B49 cDNAs. The sequences are aligned with gaps (dots) introduced to achieve maximum homology. Only sequence differences that occur in B49 and A30 are listed above and below A20. The asterisks precede the first nucleotide in the A30 and B49 sequences and follow the last nucleotide in all three sequences. Non-coding sequences are listed in groups of 10 nucleotides and coding sequences in groups of three. Boxed sequences are potential poly(A) addition signals.

the position corresponding to the amino terminus of the mature protein. Initiation of translation at only the proximal ATG is consistent with Kozak's modified scanning hypothesis (Kozak, 1981). The consensus sequence (A/G)XXAUGG derived from 153 mRNAs examined by Kozak is in agreement with the sequence at the first ATG in A20. Of the non-functional ATGs present in the 5′ non-coding region of a few eucaryotic messages, none show agreement with this consensus sequence.

The signal peptide regions of A20 and A30 can therefore be assumed to be 21 amino acids in length. Both have features present in other eucaryotic signal peptides; a basic amino acid close to the amino terminus followed by a long sequence of hydrophobic amino acid residues (Inouye and Halegoua, 1980). There are nine nucleotide differences in this region of these two clones, resulting in six amino acid changes. This extent of divergence is slightly higher than that seen for the remainder of the protein sequences, similar to the divergence seen among other related genes (Blobel *et al.*, 1979). The first 19 amino acids of the mature protein, however, are identical in A20 and A30. This is consistent with the suggestion of Burr and Burr (1981) that the tertiary structure of the region immediately following the signal peptide might be important for proper processing.

*General features of the zein sequences*

Many features of the amino acid sequence derived from the A30 clone are also present in the amino acid sequence deduced from clones A20 and B49. All three have a prevalence of dipeptide repeats (e.g., ala-ala, leu-leu, gln-gln), and the most striking feature of A30, the repeated structure of 20 amino acids at the carboxy-terminal two-thirds, is present in both. We have aligned this region of A20 and B49 (Figure 4) to emphasize this feature, with a consensus sequence shown below each. From the amino acid repeats of all three, we have drawn a combined consensus which is listed at the bottom of the figure. It is not entirely clear where the repetitive sequences begin in A20. We have included all the sequences showing significant homology, although the first repeat is interrupted by three single amino acid insertions. Regardless of where the repeat sequence begins, however, the same consensus sequence can be derived in a slightly different but equivalent form.

The repeated structure is highly conserved in A20 where 75% of the amino acids among its repeats agree with the combined consensus. A30 has a similar value of 74%. This region of B49 shows more divergence with insertions and/or deletions breaking up the regularity of the repeats to a greater extent. However, in the alignment shown in Figure 4, a great deal of similarity is apparent among the repeats: 70% of the amino acids agree with the combined consensus sequence. The relatedness of these repeats within a given clone is even more apparent at the nucleotide level. Using A20 as an example (almost exactly the same can be said of A30 or B49), at 55 of the 60 nucleotides in the repeated sequence the same base occurs at least 50% of the time and in 42 of 60 there is a two-thirds or greater chance of finding the same base. The homology at the amino acid level is closer than is first apparent. Again in A20, 23 of the 33 non-matching amino acids differ from the consensus due to a single base substitution.

This region of these zein genes has some interesting similarities to the collagen genes (Yamanda *et al.*, 1980). The results reported for the collagen genes imply that the structure was derived from an ancestral sequence by multiple duplications of a 54-nucleotide coding segment. These sequences appear to have changed relative to one another by point muta-

| | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A20 | Met | Ala | Thr | Lys | Ile | Phe | Ser | Leu | Leu | Met | Leu | Leu | Ala | Leu | Ser | Ala | Cys | Val | Ala | Asn |
| A30 | | | Ala | | | | Cys | | | | | | Gly | | | | Ser | Ala | | Thr |
| A20 | Ala | Thr | Ile | Phe | Pro | Gln | Cys | Ser | Gln | Ala | Pro | Ile | Ala | Ser | Leu | Leu | Pro | Pro | Tyr | Leu |
| A30 | | | | | | | | | | | | | | | | | | | | |
| A20 | Pro | Ser | Met | Ile | Ala | Ser | Val | Cys | Glu | Asn | Pro | Ala | Leu | Gln | Pro | Tyr | Arg | Leu | Gln | Gln |
| A30 | Ser | Pro | Ala | Val | Ser | | | | | | | Ile | | | | | | Ile | | |
| A20 | Ala | Ile | Ala | Ala | Ser | Asn | Ile | Pro | Leu | Ser | Pro | Leu | Leu | Phe | Gln | Gln | Ser | Pro | Ala | Leu |
| A30 | | | | | Gly | Ile | Leu | | | | | | Phe | Leu | | | | Ser | | |
| A20 | ... | ... | ... | ... | Ser | Leu | Val | Gln | Ser | Leu | Val | Gln | Thr | Ile | Arg | Ala | Gln | Gln | Leu | Gln |
| A30 | Leu | Gln | Gln | Leu | Pro | | | His | Leu | | Ala | | Asn | | | | | | | |
| A20 | Gln | Leu | Val | Leu | Pro | Val | Ile | Asn | Gln | Val | Ala | Leu | Ala | Asn | Leu | Ser | Pro | Tyr | Ser | Gln |
| A30 | | | | | ... | ... | ... | ... | ... | ... | ... | ... | | | | Ala | Ala | | | |
| B49 | | | | | | | | | | * | Ala | Met | Val | | | | | | | |
| A20 | Gln | Gln | Gln | Phe | Leu | Pro | Phe | Asn | Gln | Leu | Ser | Thr | Leu | Asn | Pro | Ala | Ala | Tyr | Leu | Gln |
| A30 | | | | | | | | | | | Ala | Ala | | | Ser | | Ser | | | |
| B49 | | | Gln | | Ile | Ser | Ser | | Pro | | Asp | Val | Val | Asn | | Pro | Thr | | Leu | |
| A20 | Gln | Gln | ... | Leu | Leu | Pro | Phe | Ser | Gln | Leu | ... | ... | ... | ... | Ala | Thr | Ala | Tyr | Ser | Gln |
| A30 | | | ... | Gln | | | | | | | ... | ... | ... | ... | Pro | Ala | | | Pro | |
| B49 | | | ... | | | Gln | Gln | Ile | Ile | Pro | Ala | Leu | Thr | | | | Val | Ala | | |
| A20 | Gln | Gln | Gln | Leu | Leu | Pro | ... | ... | ... | ... | ... | Phe | Asn | Gln | Leu | Ala | Ala | Leu | Asn | Pro |
| A30 | | ... | ... | Phe | | | ... | ... | ... | ... | ... | | | | | | | | | Ser |
| B49 | | | | | | | ... | ... | | | | | Asn | | | Thr | Val | Ser | | Ser |
| A20 | Ala | Ala | Tyr | Leu | Gln | Gln | Gln | Ile | Leu | Leu | Pro | Phe | Ser | Gln | Leu | Ala | Ala | Ala | Asn | Arg |
| A30 | Pro | | | | | | | Gln | | | | | | | | | Gly | Val | Ser | Pro |
| B49 | | Ala | Tyr | | Gln | | Arg | | Gln | | Leu | ... | Asn | Pro | Leu | Val | Val | | | |
| A20 | Ala | Ser | Phe | Leu | Thr | Gln | Gln | Gln | Leu | Leu | Pro | Phe | Tyr | Gln | Gln | Phe | Ala | Ala | Asn | Pro |
| A30 | | Thr | | | | | Pro | | | | | | | | His | Ala | | Pro | | Ala |
| B49 | Leu | Val | | Ala | Phe | | | Gln | | | | | | Tyr | Asn | | Phe | Ser | | Met |
| A20 | ... | ... | Ala | Thr | Leu | Leu | Gln | Leu | Gln | Gln | Leu | Leu | Pro | Phe | Val | Gln | Leu | Ala | Leu | Thr |
| A30 | ... | ... | Gly | | | | | | | | | | | | Asn | | | | | |
| B49 | Asn | | | Leu | | Trp | | | Pro | | Val | | | | Ile | | | | | |
| A20 | Asp | Pro | Ala | Ala | Ser | Tyr | Gln | Gln | His | Ile | Ile | Gly | Gly | Ala | Leu | Phe | | | | |
| A30 | Asn | Leu | | | Phe | | | | Pro | | | | | | | | | | | |

**Fig. 3.** The amino acid sequences derived from the nucleotide sequences (Figure 2) of the A20, A30, and B49 cDNAs. These sequences are aligned as in Figure 2. The asterisk precedes the first amino acid of the B49 coded protein.

tions and to a lesser extent by additions or deletions. The homology among the putative nucleotide repeats of the zein gene is even greater than among those of collagen. The zein repeat also contains some basic features of the zein proteins: amino acid composition and the prevalence of dipeptide repeats. Each collagen repeat has a substructure at the amino acid level where multiples of Gln-X-Y appear. Zein has no similar substructure although portions are repetitive.

One significant difference between these two gene families is a possible mechanism for the duplication of these repeats. Each of the collagen repeats is bounded by introns. This uniquely defines these regions of the collagen gene and implies a very plausible mechanism for the evolution of this gene from an ancestral sequence. No introns have been detected in zein genes (Wienand *et al.*, 1981; Hu *et al.*, 1982), thus the zein repeat is not so easily defined. Two additional factors argue against the idea that the repeated regions of the zeins arose by a series of duplications of 60 nucleotides. First, the insertions, deletions, and duplications seen from the comparison of these and other clones sequenced in our laboratories are of a wide variety of sizes (e.g., 96, 24, 12, 6, 3 bp). Secondly, the homology of the repeats within a gene imply at least part of this region arose by a larger duplication. This is best illustrated by the A20 sequences underlined in Figure 4. These two sequences, having 88% homology, would appear to be direct repeats of 108 nucleotides. This is significantly more homology than any other direct repeats derived from this region. This, of course, does not exclude a 60-nucleotide duplication being responsible for the evolution of other regions of these sequences. It does, however, suggest that the mechanisms of evolution of an ancestral zein sequence are likely to have been very complex.

**Table I.** Amino acid composition of zein derived from the nucleotide sequences of clones A30, A20, and B49

| | A30 | A20 | B49 |
|---|---|---|---|
| Non-polar | | | |
| Ala | 29 (14) | 30 (14) | 16 (13) |
| Ile | 9 (4) | 10 (5) | 5 (4) |
| Leu | 43 (20) | 42 (19) | 22 (18) |
| Met | 0 (0) | 1 (0.5) | 2 (2) |
| Phe | 13 (6) | 12 (5) | 4 (3) |
| Pro | 23 (11) | 22 (10) | 11 (9) |
| Trp | 0 (0) | 0 (0) | 1 (1) |
| Val | 5 (2) | 7 (3) | 9 (7) |
| Sum | 122 (57) | 124 (57) | 70 (57) |
| Polar | | | |
| Asn | 10 (5) | 10 (5) | 9 (7) |
| Cys | 2 (1) | 2 (1) | 0 (0) |
| Gln | 41 (19) | 43 (20) | 24 (20) |
| Gly | 5 (2) | 2 (1) | 2 (2) |
| Ser | 15 (7) | 17 (8) | 7 (6) |
| Thr | 5 (2) | 7 (3) | 3 (2) |
| Tyr | 8 (4) | 8 (4) | 5 (4) |
| Sum | 86 (40) | 89 (41) | 50 (41) |
| Basic | | | |
| Arg | 2 (1) | 3 (1) | 1 (1) |
| His | 2 (1) | 1 (0.5) | 0 (0) |
| Lys | 0 (0) | 0 (0) | 0 (0) |
| Sum | 4 (2) | 4 (2) | 1 (1) |
| Acidic | | | |
| Asp | 0 (0) | 1 (0.5) | 1 (1) |
| Glu | 1 (0.5) | 1 (0.5) | 0 (0) |
| Sum | 1 (0.5) | 2 (1) | 1 (1) |

Numbers in parenthesis indicate the percentage of the total number of residues. These numbers do not include the signal peptide of A20 and A30.

*5′ Sequence of A20.* Upon initial examination of the 5′ non-coding region of the A20 sequence a rather bizarre feature presents itself. The first 47 nucleotides are an inverted repeat of the terminal 47 nucleotides at the 3′ end. However, when compared to Z4, a genomic clone related to the A30 subfamily (Hu *et al.*, 1982) no significant homology is seen from positions 1 – 37, while the remainder of A20 and Z4 have nearly 80% homology. A similar repeat has been seen in a bovine parathyroid hormone clone and is thought to be an artifact of cDNA synthesis (Weaver *et al.*, 1981). It seems likely that this 5′ region of the A20 sequence also resulted from the cDNA 'looping back' on itself during reverse transcription and is not present in the mRNA.

Unfortunately, this complication makes it difficult to define the beginning of this zein mRNA. We have also sequenced an A20-like cDNA clone which has a 5′ end at what corresponds to position 44 in A20 (Heidecker *et al.*, in preparation). Also, while nucleotides 38 – 47 of A20 are an inversion of the corresponding region in the 3′ end, this region of homology would be expected to exist in the mRNA if the larger inverted repeat did indeed result from a mechanism analogous to the models proposed (Weaver *et al.*, 1981). We will assume then, for the purposes of discussion, that nucleotide 44 is the start of the 5′ non-coding region in A20. The corresponding position in the genomic clone Z4 is 29 bp downstream from the TATA box, implying by analogy to other gene structures (Efstratiadis *et al.*, 1980) that the A20 clone is nearly a complete copy of its message.

As might be expected, no significant homology is seen when this region is compared to other known leader sequences. Its length (≅68 bp), however, is similar to other mRNAs (Ingolia and Craig, 1981). One noteworthy difference is the high percentage of adenine (47%). While unusual, this feature is seen in the leader sequences of the *Drosophila melanogaster* heat-shock mRNAs where the A content ranges from 46% to 51% (Ingolia and Craig, 1981). Among the few plant genes sequenced, the leghemoglobin gene of soybean (Hyldig-Nielsen *et al.*, 1982) show upon examination to have 57 – 69% adenine in this region depending on where the transcript begins. A second interesting feature of this region is the sequence AGCAACA which appears three times (four if one base change is allowed). This sequence also appears as part of the larger repeat in the coding region discussed above.

*3′ Non-coding sequences.* A comparison of the nucleotide sequences of A20 and A30 in the 3′ non-coding region shows them to be of identical length (87 bp). There are 14 nucleotide substitutions representing about the same extent of divergence seen in the coding region. The two longest stretches of homology are 20 and 11 nucleotides and contain sequences similar to polyadenylation signals. By contrast, this region of B49 has a length of 97 bp and much less homology (53%), to either of the A20 or A30 sequences. Three deletions/insertions are required to achieve the alignment shown,

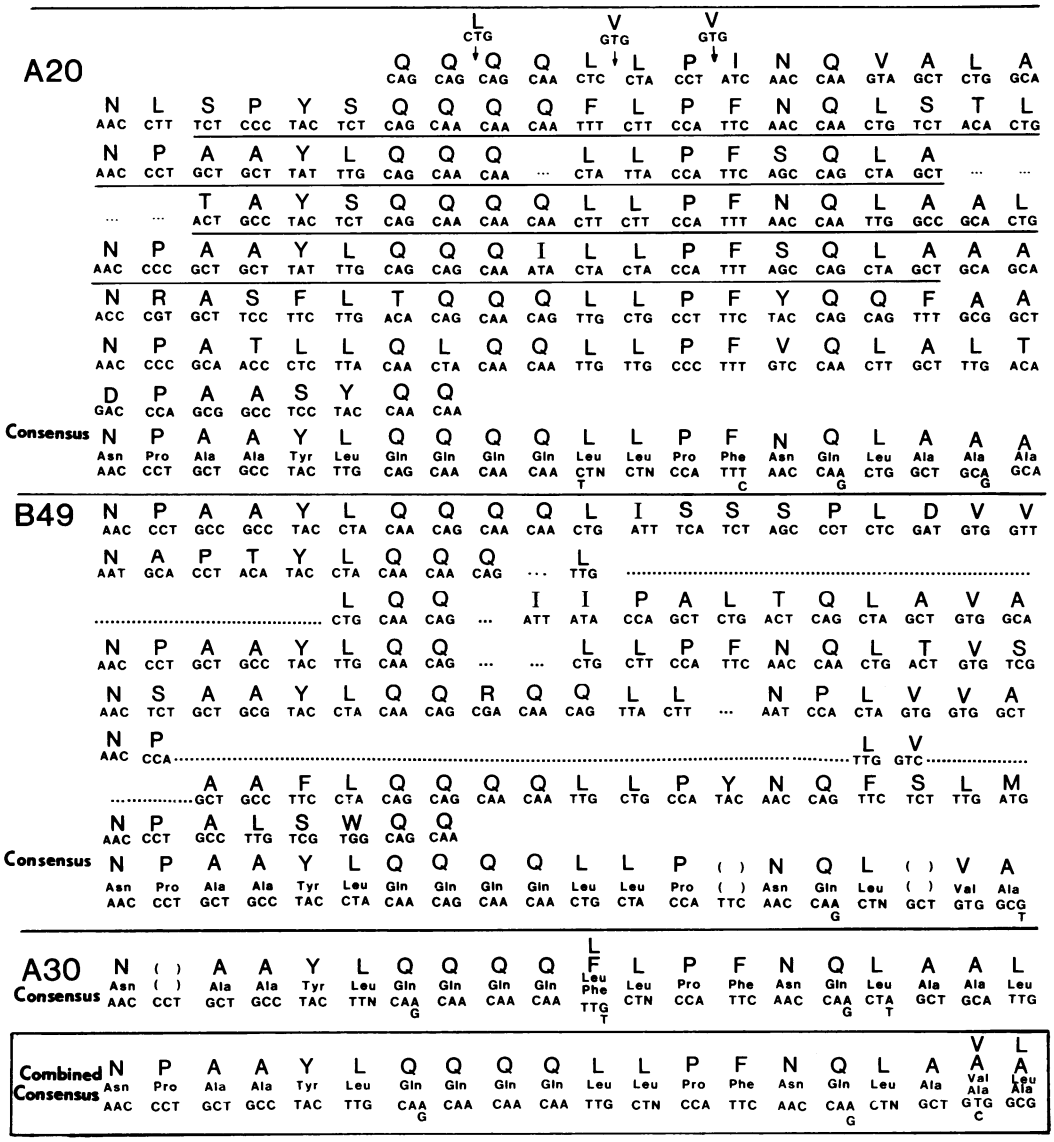


**Fig. 4.** The repetitive nucleotide and deduced amino acid sequences of the zein cDNAs. The alignment serves to maximize homology among the repeats. A consensus sequence derived from A20 and B49 is listed below each and a consensus sequence from A30 is included for comparison.

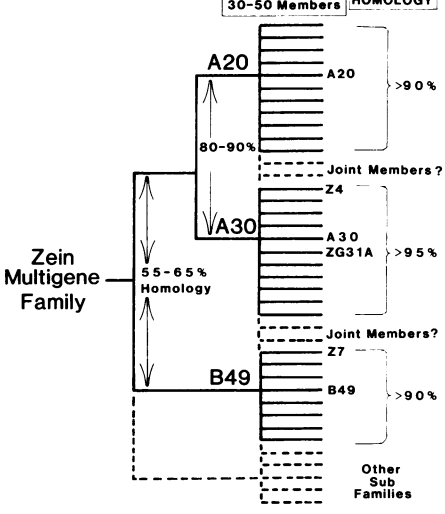


**Fig. 5.** The sequence relationship among the zein subfamilies.

and the longest region of homology (8 bp) also contains a possible polyadenylation signal.

The polyadenylation signal common to most animal mRNAs is the hexanucleotide AATAAA 12–33 bases upstream from the apparent start of the poly(A) tail (Benoist *et al.*, 1980). Of the plant mRNA sequences published, at least one has this signal (Croy *et al.*, 1982) while a variant was noted in A30. The same variant is seen in A20 and a possible second variant is seen in B49 (Figure 2). A third hexanucleotide (GATAAA) is seen in the leghemoglobin genes (Hyldig-Nielsen *et al.*, 1982). Another feature common to the nucleotide sequences of A20, A30, and the leghemoglobin genes, is a normal signal for polyadenylation appearing 60–90 nucleotides preceding the start of the poly(A) tail. It is possible that this sequence has some involvement in processing these mRNAs, although it is not present in B49. Apparently the position and/or sequence of the usual signal for polyadenylation seen in animal mRNAs is more variable in these plant mRNAs.

*Codon usage.* As reported previously for A30 and seen in

**Table II.** Codon usage in clones A30, A20, and B49

| | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Phe | TTT | 4 | 8 | 1 | Ser | TCT | 6 | 6 | 3 | His | CAT | 1 | 0 | 0 | Pro | CCT | 7 | 6 | 5 |
| | TTC | 10 | 5 | 3 | | TCC | 1 | 4 | 0 | | CAC | 1 | 1 | 0 | | CCC | 5 | 7 | 1 |
| Leu | TTA | 7 | 3 | 1 | | TCA | 5 | 5 | 1 | Gln | CAA | 28 | 27 | 13 | | CCA | 8 | 9 | 5 |
| | TTG | 11 | 12 | 6 | | TCG | 1 | 1 | 2 | | CAG | 13 | 16 | 11 | | CCG | 3 | 0 | 0 |
| | CTT | 11 | 12 | 2 | | AGT | 1 | 0 | 0 | Asn | AAT | 0 | 0 | 2 | Thr | ACT | 0 | 1 | 2 |
| | CTC | 6 | 6 | 1 | | AGC | 3 | 3 | 1 | | AAC | 10 | 11 | 7 | | ACC | 3 | 3 | 0 |
| | CTA | 10 | 8 | 6 | Arg | AGA | 0 | 0 | 0 | Lys | AAA | 1 | 0 | 0 | | ACA | 2 | 4 | 1 |
| | CTG | 3 | 6 | 6 | | AGG | 2 | 2 | 0 | | AAG | 0 | 1 | 0 | | ACG | 1 | 0 | 0 |
| Ile | ATT | 3 | 2 | 2 | | CGT | 0 | 1 | 0 | Asp | GAT | 0 | 0 | 1 | Ala | GCT | 15 | 16 | 6 |
| | ATC | 5 | 4 | 2 | | CGC | 0 | 0 | 0 | | GAC | 0 | 1 | 0 | | GCC | 5 | 5 | 7 |
| | ATA | 2 | 5 | 1 | | CGA | 0 | 0 | 1 | Glu | GAA | 1 | 1 | 0 | | GCA | 11 | 10 | 2 |
| Met | ATG | 2 | 3 | 2 | | CGG | 0 | 0 | 0 | | GAG | 0 | 0 | 0 | | GCG | 4 | 4 | 1 |
| Val | GTT | 0 | 1 | 2 | Tyr | TAT | 2 | 3 | 0 | Gly | GGT | 4 | 2 | 1 | Trp | TGG | 0 | 0 | 1 |
| | GTC | 0 | 1 | 1 | | TAC | 6 | 5 | 5 | | GGC | 2 | 0 | 0 | Stop | TAA | 0 | 0 | 0 |
| | GTA | 1 | 3 | 0 | Cys | TGT | 1 | 2 | 0 | | GGA | 0 | 0 | 1 | | TAG | 1 | 1 | 1 |
| | GTG | 4 | 3 | 6 | | TGC | 2 | 1 | 0 | | GGG | 0 | 0 | 0 | | TGA | 0 | 0 | 0 |

The numbers refer to codon usage in A30 (first column), A20 (second column), and B49 (third column), respectively.

most other messages sequenced, codon usage in A20 and B49 is non-random. Most of the codon preferences previously reported for the A30 sequence are also present in the latter two clones. The dichotomy noted for A30 in the use of the asparagine codons is also present in A20 (see Table II). This is true even though five of 11 uses of this codon in A20 do not simply reflect sequence homology with A30. A20 and A30 have a preference for CAA over CAG in coding glutamine of ~2:1. Also CTT and TTG are favored in coding leucine. When codon usage in B49 is compared with the terminal one-half of A20 and A30 it, too, is similar. Thus, while CTT is used twice in B49, it is used three times in the corresponding region of A20.

Grantham *et al.* (1981) have compiled tables of codon usage for most of the sequenced mRNAs. Few plant sequences are known and it is of interest to see how codon usage among the sequences presented here compares with that of animals. The same avoidance of the CG doublet is seen, and many of the asymmetries seen in their compilation are very much like those seen here (e.g., Pro, Asn, Phe). Some striking differences are also present and are at least partly related to the preference of G over A in the third position. Whereas animal mRNAs contain more degenerate G than A, some of the cases here show nearly the opposite preference. This is illustrated by glutamine where the codon preference in animals shows three times as much degenerate G while, as noted above, the zein clones have twice as much degenerate A. Leucine also differs, with CTC and CTG being strongly preferred in animals. Here they are among the least frequently used. Future comparisons with other plant mRNAs will be needed to show if these observations represent a general feature or are unique to the zein sequences.

## Conclusion

The similarities among these zein cDNAs, along with other characterizations of zein clones (Park *et al.*, 1980; Burr *et al.*, 1982), are consistent with the idea that the members of the zein gene family have evolved from a common ancestral sequence. The divergence of these genes has taken place through a variety of duplications, insertions, deletions, and nucleotide substitutions, all of which have conserved both the amino acid composition and the repeated structure of the zeins. This relationship, among what we have termed the subfamilies of zeins, is illustrated in Figure 5. Each subfamily so far identified consists of an undetermined number of members and is defined by sequence homology to one of the three clones discussed in this paper. Burr *et al.* (1982) have also characterized at least one additional clone (B59) which is apparently a member of another distinct subfamily. We have also allowed for the possibility that some zein sequences are members of two distinct subfamilies. The results of Viotti *et al.* (1982) suggest that this may indeed be the case.

This grouping is more appropriate than the normally used classification based on mol. wt. While A30 binds mRNAs which translate into 19 000 mol. wt. zeins, a small amount of 22 000 mol. wt. zein can also be seen (Park *et al.*, 1980). Consistent with this is the finding that Z4, while having 97% homology to A30, encodes a protein corresponding to a 22 000 mol. wt. zein due to a duplication in the middle of the coding region (Hu *et al.*, 1982). It is possible that the subfamily defined by B49 also contains lower mol. wt. zeins. This classification might also be useful in understanding mutations affecting zein synthesis. Some mutations thought to be involved in regulating zein synthesis affect one of the mol. wt. classes to a greater extent (Soave *et al.*, 1978). It will be of interest to see if these mutations act specifically on one or more of the zein subfamilies.

## Materials and methods

*Plasmids and cloning vectors*

The plasmids A20 and B49 and their isolation and characterization have been described by Burr *et al.* (1982) and Park *et al.* (1980). The phage vector M13mp7 and its growth conditions have been described by Messing *et al.* (1981).

*Sequence analysis*

Restriction fragments were end-labeled, and either cut with second restriction endonuclease or strand separated and sequenced as described by Maxam and Gilbert (1980). The A + G reaction was modified in a manner similar to that described by Bernard and Gough (1980). 10 μl labeled DNA were incubated with 15 μl 100% formic acid at 20°C for 3 – 10 min. The reaction was stopped by the addition of 200 μl hydrazine stop solution and ethanol precipitated.

Dideoxy sequencing with the phage M13mp7 was as previously described in Sanger *et al.* (1977) and Messing *et al.* (1981).

*Chemicals and enzymes*

Restriction enzymes, T4 DNA ligase, and polynucleotide kinase were purchased from Bethesda Research Labs or New England Biolabs and used as recommended by the supplier. [$\gamma$-$^{32}$P]ATP was obtained from ICN and [$\alpha$-$^{32}$P]dATP was obtained from Amersham.

*Computer analysis*

All sequence data were entered and stored on computer diskettes and analyzed in part with the programs developed by Larson and Messing (1982) on an Apple II plus computer.

## Acknowledgements

We thank Kae Ebling and Kris Kohn for help in preparing this manuscript. We also acknowledge the technical assistance of Mark Peifer in the early part of this work. 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-ACO2-81ER 10901; from the Minnesota Agriculture Experiment Station, MN-15-030; and from NRSA Grant GM07094.

## References

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

Bernard,O., and Gough,N.M. (1980) *Proc. Natl. Acad. Sci. USA,* **77**, 3630-3634.

Blobel,G., Walter,P., Chang,C.N., Goldman,B.M., Erickson,A.H., and Lingappa,V.R. (1979) in Hopkins,C.R., and Duncan,C.J. (eds.), *Secretory Mechanisms S.E.B., Symp. 33,* Cambridge University Press, pp. 9-36.

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

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

Burr,F.A., and Burr,B. (1981) in Redei,G.P. (ed.,), *Stadler Genetics Symp.* **13**, University of Missouri Press, pp. 79-92.

Croy,R.R.D., Lycett,G.W., Gatehouse,J.A., Yarwood,J.N., and Boulter,D. (1982) *Nature,* **295**, 76-79.

Efstratiadis,A., Posakony,J.W., Maniatis,T., Lawn,R.M., O'Connell,C., Spritz,R.A., DeRiel,S.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.

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

Grantham,R., Gautier,C., Gouy,M., Jacobzone,M., and Mercier,R. (1981) *Nucleic Acids Res.,* **9**, r43-r74.

Hagen,G., and Rubenstein,I. (1980) *Plant Sci. Lett.,* **19**, 217-233.

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

Hu,N.-T., Peifer,M., Heidecker,G., Messing,J., and Rubenstein,I. (1982) *EMBO J.,* **1**, 1337-1342.

Hyldig-Nielsen,J.J., Jensen,E.O., Paludan,K., Wilborg,O., Garret,R., Jorgensen,P., and Marcker,K.A. (1982) *Nucleic Acids Res.,* **10**, 689-701.

Ingolia,T.D., and Craig,E.A. (1981) *Nucleic Acids Res.,* **9**, 1627-1642.

Inouye,M., and Halegoua,S. (1980) *CRC Crit. Rev. Biochem.,* **7**, 339-371.

Kozak,M. (1981) *Nucleic Acids Res.,* **9**, 5233-5251.

Larkins,B.A., and Hurkman,W.J. (1978) *Plant Physiol.,* **62**, 256-263.

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

Maxam,A.M., and Gilbert,W. (1980) in Grossman,L., and Moldave,K. (eds.), *Methods in Enzymology,* Vol. **651**, Academic Press, NY, pp. 499-560.

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

Misra,P.S., Mertz,E.T., and Glover,D.V. (1975) in CYMMIT-Purdue University (ed.), *High Quality Protein Maize,* Dowden, Huchinson and Ross, pp. 291-305.

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

Pederson,K., Bloom,K.S., Anderson,J.N., Glover,D.V., and Larkins,B.A. (1980) *Biochemistry (Wash.),* **19**, 1644-1650.

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

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

Soave,C., Righetti,P.G., Lorenzoni,C., Gentinetta,E., and Salamini,F. (1976) *Maydica,* **21**, 61-75.

Soave,C., Suman,N., Viotti,A., and Salamini,F. (1978) *Theor. Appl. Genet.,* **52**, 263-264.

Soave,C., Reggiani,R., DiFonzo,N., and Salamini,F. (1981) *Genetics,* **97**, 363-377.

Soave,C., Reggiani,R., DiFonzo,N., and Salamini,F. (1982) *Biochem. Genet.,* in press.

Viotti,A., Sala,E., Marotta,R., Alberi,P., Balducci,C., and Soave,C. (1979) *Eur. J. Biochem.,* **102**, 213-222.

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

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

Weaver,C.A., Gordon,D.F., and Kemper,B. (1981) *Proc. Natl. Acad. Sci. USA,* **78**, 4073-4077.

Wienand,U., and Feix,G. (1980) *FEBS Lett.,* **116**, 14-16.

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

Yamanda,Y., Avvedimento,V.E., Mudryi,M., Ohkubo,H., Vogeli,G., Irani, M., Pastan,I., and deCrombrugghe,B. (1980) *Cell,* **22**, 887-892.