Mol Gen Genet (1989) 216:81–90

MGG

Molecular characterization of an active wheat LMW glutenin gene and its relation to other wheat and barley prolamin genes

Vincent Colot, Dorothea Bartels*, Richard Thompson*, and Richard Flavell

Department of Molecular Genetics, AFRC Institute of Plant Science Research (Cambridge Laboratory),
Trumpington, Cambridge CB2 2JB, UK

Summary. The isolation and characterisation by DNA sequencing of a low molecular weight (LMW) glutenin gene from wheat is described. The deduced protein contains a signal peptide, a central repetitive region rich in proline and glutamine and N and C terminal non-repetitive domains, similar to other prolamins. A detailed comparison of the C terminal domain of 20 prolamin genes enabled us to divide them into 4 families. The LMW glutenin family is distinct from the α, β- and γ-gliadin families of wheat and is closest to the B hordein genes of barley. This and other comparisons were also used to assess the pattern of genetic variation among prolamin sequences and to provide a molecular basis for the interpretation of prolamin size polymorphism. The 5′ flanking fragment of the isolated gene was previously shown to direct endosperm-specific expression of a reporter gene in transgenic tobacco. Evidence is provided that the isolated gene is also active in wheat and its transcription initiation site was determined. Features of the gene which may be relevant to its activity are discussed.

Key words: LMW glutenin – Seed storage protein – Molecular evolution – Tissue-specific expression – Wheat

Introduction

The major storage proteins of hexaploid wheat have long been recognised for their important role in determining the nutritive and baking properties of flour. These proteins are called prolamins because of their high proline and glutamine content. Prolamins are specifically synthesized and deposited in the endosperm of the developing grain, where they constitute the primary source of nitrogen for the onset of protein synthesis that occurs during subsequent germination and early growth.

Wheat prolamins have been classified into two groups, the glutenins and the gliadins, according to their solubility in aqueous solvents, and have been further categorized according to their size into five related protein families (α,β-, γ- and ω-gliadins; LMW and HMW glutenins, reviewed in Payne 1987). Genes encoding members of three wheat prolamin families, the HMW glutenin, the α,β-gliadin and the γ-gliadin families have now been described (Okita et al. 1985; Sumner-Smith et al. 1985; Thompson et al. 1985; Bartels et al. 1986; Rafalski 1986; Halford et al. 1987). Although no ω-gliadin gene has yet been characterized, a related C-hordein has recently been sequenced (Entwistle 1988). Finally, several cDNA clones have been isolated that may encode LMW glutenin components (Bartels and Thompson 1983; Okita et al. 1985). As both wheat γ-gliadins and LMW glutenins are encoded by genes clustered at the complex *gli-1* locus (Payne et al. 1984), the relationship of these genes needs to be established unequivocally.

A partial sequence of a LMW glutenin gene 5′ flanking region that was used in transgenic tobacco experiments was presented previously (Colot et al. 1987). In this paper we describe the nucleotide sequence of the entire gene (now called *LMWG-1D1*). Southern blot analysis shows that the gene *LMWG-1D1* resides on chromosome 1D and is one of a small family carried on group 1 chromosomes of hexaploid wheat.

Comparisons of *LMWG-1D1* with other known prolamin gene sequences of wheat and barley, indicate that the isolated gene belongs, together with sequences originally defined by Okita et al. (1985) as γ-type gliadins, to a gene family which encodes LMW glutenin subunits and is markedly distinct from the γ-gliadin gene family defined by Bartels et al. (1986) as well as from α,β-gliadin genes. Furthermore, *LMWG-1D1* is closely related to barley B1 and B3 hordein sequences (Forde et al. 1985a, b; Brandt et al. 1985). Comparisons between several prolamin gene sequences are described in this paper in order to assess the pattern of genetic variation among prolamin sequences and to provide a molecular basis for the size polymorphism which is observed at the protein level both within and between prolamin families.

S1 mapping and primer extension experiments have been performed which show that *LMWG-1D1* is active in wheat endosperm. Sequences similar to upstream elements important for gene expression in a variety of eukaryotes, including v-JUN/GCN4 binding sites, have been found in the 5′ flanking region of the wheat gene described here. With respect to the 3′ flanking region, different polyadenylation sites are used by different members of the LMW glutenin gene family defined here.

* *Present address:* Max Planck Institut für Züchtungsforschung, Abteilung Salamini, Egelspfad 1, D-5000 Köln 30, Federal Republic of Germany

Offprint requests to: V. Colot

Materials and methods

Plant material. The hexaploid wheat variety Chinese Spring and the various nullisomic-tetrasomic lines derived from it by Sears (1966) were obtained from the wheat collection maintained by Dr. Law and colleagues at the PBI.

Developing endosperms were isolated from ears of field-grown wheat at 2–3 weeks post-anthesis.

DNA and RNA extraction. High molecular weight wheat DNA was prepared from 4-day-old etiolated shoot tips, as described previously (Thompson et al. 1983, method 1).

Endosperm poly(A)$^+$ mRNA was extracted according to Bartels and Thompson (1983).

Bacterial strains and bacteriophages. Escherichia coli strain WL 268 and lambda vector Charon 35 were provided by Dr. Loenen (Loenen and Blattner 1983), and *E. coli* strain K 803 by Dr. Fedoroff (Fedoroff 1983).

*Genomic library construction. Bam*H1 partial digestion products of wheat DNA in the 10–25 kb range were obtained as described in Maniatis et al. (1982). The lambda Charon 35 vector was prepared by first ligating the cohesive ends and then digesting with *Bam*H1. The ligated arms were separated from the "stuffer" fragments by sucrose gradient fractionation. The size-fractionated wheat DNA, partially cleaved with *Bam*H1, was ligated into the *Bam*H1 sites of the lambda vector and the mixture was packaged in vitro. The packaged mixture was plated on *E. coli* strain K803 (RecA$^+$), a strain partially tolerant of the high degree of CpG methylation found in wheat DNA (see Fedoroff 1983).

Clone identification and purification. The phage library was screened by plaque hybridization using as a probe the insert of the LMW glutenin cDNA clone pTag544 (Bartels and Thompson 1983). Hybridizing plaques were picked and purified by several rounds of plating with *E. coli* strain WL268 (RecA$^-$), as discussed in Thompson et al. (1985). Lambda clone DNA was prepared as described by Cameron et al. (1977), following growth of phage on plate lysates, and the restriction sites of several enzymes were mapped. Chosen DNA restriction fragments were subcloned in the plasmid pUC19.

DNA sequencing. The gene sequence was obtained on both strands after generating unidirectional deletion by exonuclease III and S1 nuclease treatment of the appropriate insert in M13mp18 (Henikoff 1984), using the dideoxy chain termination method of Sanger et al. (1977) as modified by Messing (1983) for M13 sequencing. The sequence was completed after additional subcloning of chosen restriction fragments. The Staden computer programmes were used for sequence handling and primary analysis (Staden 1982).

S1 nuclease mapping and primer extension. The preparation of a radioactively labelled probe for experiments using S1 nuclease protection of RNA was done by the M13 second strand synthesis method of Nasmyth (1983). The probe was selected from a set of deletions made during the sequencing of the gene.

The synthetic 24-mer primer (5′-ACCAGGGATGCATCTAGTCTCCAT-3′) was chosen to be complementary to the sequence encoding amino acids 21–29 of the putative LMW glutenin protein specified by the cloned gene fragment. The oligonucleotide was labelled at the 5′ end with γ-[^{32}P]ATP using T4 polynucleotide kinase, before hybridization to mRNA.

Computer analysis. The programmes DOTPLOT, GAP, PRETTY and LINEUP of the University of Wisconsin Genetics Computer Group were used as described in the text for nucleic acid and protein sequence comparison (Devereux et al. 1984).

Results

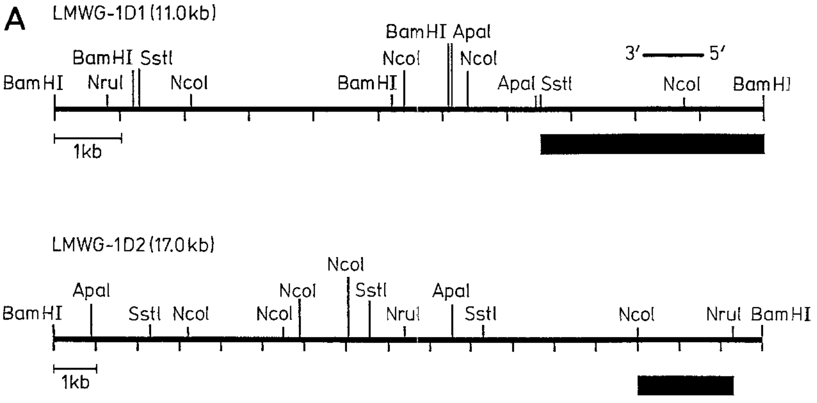
Isolation of LMW glutenin genomic clones

To define the LMW glutenin gene family of wheat prolamins, we isolated lambda clones that hybridize under high stringency (0.075 M Na$^+$, 65° C) to the putative LMW glutenin cDNA pTag544 but not to the γ-gliadin cDNA pTag647 (Bartels et al. 1986). The cleavage sites for several restriction endonucleases were mapped on the inserts of the two clones that hybridized most strongly, LMWG-1D1 and LMWG-1D2 (Fig. 1A). Each was found to contain only one *Bam*H1 + *Sst*1 subfragment hybridizing to either pTag544 or to endosperm poly(A)$^+$ RNA. This and further restriction mapping suggested that only one prolamin gene resides within their 11 kb and 17 kb long sequences, respectively (Fig. 1A). Since cloned prolamin genes are subject to rearrangements (Thompson et al. 1985), a reconstruction experiment with Chinese Spring DNA was performed which enabled us to verify, within the limits of electrophoretic resolution, that both cloned inserts correspond to chromosomal fragments (Fig. 1B). In order to assign the chromosomal location of the genomic sequences corresponding to both inserts, the reconstruction experiment was extended to DNA from the group 1 nullisomic-tetrasomic lines of Chinese Spring wheat. The results indicate that the 3.5 kb and 6.5 kb fragments of LMWG-1D1 and LMWG-1D2, respectively, are absent from the DNA from nullisomic 1D tetrasomic 1A. Thus these genes are present on chromosome 1D, probably in 1 or 2 copies per chromosome (Fig. 1B). As also seen in Fig. 1B, there are at least 14 *Bam*H1 + *Sst*1 DNA fragments related to pTag544 in the genome of hexaploid wheat, which are similarly all located on the group 1 chromosomes.

As a first step towards the characterisation of this gene family, the DNA sequence of most of the 3.5 kb *Bam*H1-Sst1 fragment of LMWG-1D1 was determined.

Nucleotide sequence of gene LMWG-1D1

Coding region. A single large open reading frame (ORF) is found within the 3.5 kb subfragment of clone LMWG-1D1, starting 984 nucleotides downstream of the *Bam*H1 site and with a coding capacity of 307 amino acids (Fig. 2, position +45 to +965). That the ORF is potentially translated is suggested by the presence of the sequence motif CCACC$\underline{\text{ATG}}$A around the initiator codon, which closely resembles the consensus sequence motif of eukaryotic initiation sites (Kozak 1986). The ORF and the 5′ and 3′ untranslated regions of *LMWG-1D1* are uninterrupted by introns, as indicated by comparison with the closely related sequences of the full-length cDNA clone B11–33 and of the partial cDNA pTag544 (Okita et al. 1985; Bartels and Thompson 1983; data not shown).



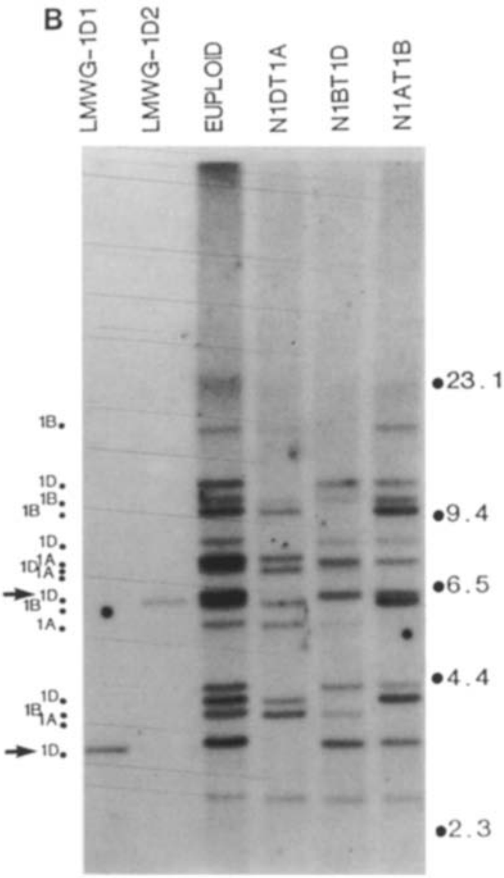


Fig. 1A and B. Restriction endonuclease maps and chromosomal localization of clones LMWG-1D1 and LMWG-1D2. **A** Restriction endonuclease maps. The fragments hybridizing to total poly(A)$^+$RNA from wheat endosperm are symbolized by the *thick bars* shown below each map. The orientation and the length of *LMWG-1D1* coding sequence is shown above the map. **B** Southern blot analysis of *Bam*H1 + *Sst*1 digests of DNA of clones LMWG-1D1 and LMWG-1D2, and of euploid and group 1 nullisomic-tetrasomic lines of Chinese Spring. Wheat DNA (7.5 µg) was loaded on the gel as well as 20 pg of cloned DNA (equivalent to 1 or 2 copies of a 5 kb long wheat DNA fragment). The DNA was probed after transfer to nitrocellulose filters with the central *Alu*1 fragment of pTag544 (Bartels and Thompson 1983; Harberd et al. 1985). LMWG-1D1 and LMWG-D2 are each located on chromosome 1D (*arrows*). Size markers in kb are indicated on the right of the figure

The predicted LMWG-1D1 protein consists of an N-terminal sequence rich in hydrophobic residues that is typical of signal peptides (von Heijne 1985), closely followed by repeats rich in proline and glutamine residues and a C-domain that is essentially a non-repeated sequence interspersed with 8 stretches of 3 or more glutamine residues (Fig. 2).

5' flanking region. To establish whether gene *LMWG-1D1* is transcribed, a hybridization probe was used which extended from positions −497 (*Aat*II site) to +139 in the sequence. The probe was hybridized to total poly(A)$^+$ RNA isolated from developing endosperm (10–20 days post-anthesis). Several different conditions were used for the S1 nuclease treatment of the mRNA probe hybrids which all resulted in a multiplicity of protected fragments, as shown in Fig. 3A. These multiple fragments can, in part, be explained by melting of the hybrid ends but may also originate from the probe hybridizing to several closely related mRNA species. However, irrespective of the conditions used, the same predominant fragment was protected from S1 nuclease digestion. Thus this fragment must correspond to an mRNA species transcribed from the *LMWG-1D1* gene or from a gene almost identical to it. The potential transcription start site (between positions −2 and 0 in Fig. 2) of *LMWG-1D1* was then deduced from a sizing of the major protected fragment.

Primer extension was also used to determine the 5' end of the mRNA. Total poly(A)$^+$ RNA from wheat endosperm was also hybridized to a synthetic oligonucleotide which served subsequently as a primer for RNA reverse transcription. The 24-mer oligonucleotide is complementary to a sequence of clone LMWG-1D1 that is unrelated to sequences of other prolamin gene families (data not shown), and corresponds to the putative signal peptide/mature protein border (positions +105 and +128 in Fig. 2). The results of the primer extension experiment are presented in Fig. 3B which show two major products, 128 and 130 nucleotides long. These product lengths designate a transcription start site for *LMWG-1D1* consistent (+ or −5 bp) with that derived from S1 nuclease mapping (Fig. 2). Thus, as also supported by our experiments in transgenic tobacco (Colot et al. 1987), we assume that *LMWG-1D1* corresponds to a transcribed gene in wheat endosperm.

Two TATA sequence motifs are found in the vicinity of the transcription start site, 14 and 31 nucleotides 5' of position 0 (Fig. 2). Their relative locations suggest that only the distal TATA sequence motif is functional (Kovacs and Butterworth 1986 and references therein). Several other motifs similar to known regulatory signals of eukaryotic gene expression are present further 5' of the gene sequence. A CCAAT box is found 120 bp upstream of the transcription start site and many sequences (not shown) similar to sequence motifs of the SV40 enhancer (Ondek et al. 1988) are scattered in both orientations, along the 1 kb 5' flanking region. Also, a DNA sequence identical to the consensus binding site of the functionally equivalent mammalian and yeast regulatory proteins v-JUN and GCN4 (Struhl 1987; 1988) is present twice, around positions −240 and −515. Apart from the motifs mentioned above and which have been frequently found in eukaryotic gene expression systems, the 5' flanking sequence of *LMWG-1D1* contains sev-

```
   -930                -910                -890                -870                -850                -830
GATCCGGGAGAGGCTGCGGCGGCGAGCGGTCCGGCCGGCGAGGAGGACGGGGAAGGGAGAGACGAAGACGACGCAGAATCGGCGGAATGTGGGCTGGGCTTGTGACTTTGAGGCCTCCGA

   -810                -790                -770                -750                -730                -710
AAACTATAGCCCAGTTCGAATTGGTGCCCTAACACACACAACACTTACGTTGGGCCTAATCGCTCGCTCCTGCCCCTGCTCAAAATTTTTTTTGCTCCAGGCTGGGGGCCTCTGTTCCAC

   -690 IV             -670           III -650                -630     II         -610                -590  I
CCCCATCTATCGCTCCACCTCCAAACAAAAAAAAAATCTATCACTCCACCTCCACTCCAAAAATATAAAATTCTATCAATCCACCTACGCCTCGAAAAAAGAAATCTATCACTCCACCTC

   -570                -550                -530                -510                -490                -470
AGCATTGATGTCTCTAGCTTGTAGAAACTGCCATCCTTTACATGTAAAACGGATTCGATGAGTCATGTCATGCTCTATAGACGTCAGTTCATCTTATCATCTTACAGGAAAGTACAAAGT
                                             ENDOSPERM BOX II
   -450                -430                -410                -390                -370                -350
TAGTTTTCTGAAAAGCAACCGAATATAGAAGAACACTCCACACTCAAGGCTTTACTAATCGAGCATATCCTAACAGCCCACACATGATTGCAAACTTAGTCATACACAAGTTTTGCCTTT

   -330                -310                -290                -270                -250                -230
CTTGTTTACGGCTGACAGCCTATACAAGGTTCCAAACTCGGTTGTAAAAGTGATACTATCTTGATAAGTGTGTGACATGTAAAGTTAATAAGGTGAGTCATATATAGCAAATATCGGGGT
                                                                                  ENDOSPERM BOX I
   -210                -190                -170                -150                -130                -110
TTCTGTACTTTGTGTGTGATCGTATGCACAACTAAAAATCAACTTTGATGATCAATATATCCAAAAGTACGCTTGTAGCTAGTGCAAACCTAACCCAATGTAACAAAATAATTCATTTCA

   -90                 -70                 -50                 -30                 -10                  10
GATGGAGCCAAACAGAATTATTAAAGCTGATGCAAAGAAGGAAAAGAGGTGGTTCCTGGGCTACTATAAATAGGCATGAAGTATAAAGATCATCACAAGCACAAGCATCAGAACCAAGCA


   30                  50                  70                  90                  110                 130
ACACTAGTTAACACCAATCCACCATGAAGACCTTCCTCGTCTTTGCCCTCCTCGCCGTTGCGGCGACAAGTGCAATTGCGCAGATGGAGACTAGATGCATCCCTGGTTTGGAGAGACCAT
                        M  K  T  F  L  V  F  A  L  L  A  V  A  A  T  S  A  I  A  Q  M  E  T  R  C  I  P  G  L  E  R  P  W
                                             SIGNAL PEPTIDE
   150                 170                 190                 210                 230                 250
GGCAGCAGCAACCATTACCACCACAACAGACATTTCCACAACAACCACTATTTTCACAACAACAACAACAACAACTATTTCCTCAACAACCATCATTTTCGCAGCAACAACCACCATTTT
 Q  Q  Q  P  L  P  P  Q  Q  T  F  P  Q  Q  P  L  F  S  Q  Q  Q  Q  Q  Q  L  F  P  Q  Q  P  S  F  S  Q  Q  Q  P  P  F  W
      START OF REPEAT DOMAIN
   270                 290                 310                 330                 350                 370
GGCAGCAACAACCACCATTTTCTCAGCAACAACCAATTCTACCACAGCAACCACCATTTTCGCAGCAACAACAACTAGTTCTACCGCAACAACCACCATTTTCACAGCAACAACAACCAG
 Q  Q  Q  P  P  F  S  Q  Q  Q  P  I  L  P  Q  Q  P  P  F  S  Q  Q  Q  Q  L  V  L  P  Q  Q  P  P  F  S  Q  Q  Q  Q  P  V

   390                 410                 430                 450                 470                 490
TTTTACCTCCACAACAATCACCTTTTCCACAACAACAACAACAACACCAACAGCTGGTGCAACAACAAATCCCTGTTGTTCAGCCATCCATTTTGCAGCAGCTAAACCCATGCAAGGTAT
 L  P  P  Q  Q  S  P  F  P  Q  Q  Q  Q  Q  H  Q  Q  L  V  Q  Q  Q  I  P  V  V  Q  P  S  I  L  Q  Q  L  N  P  C  K  V  F
END OF REPEAT DOMAIN             START OF C-DOMAIN
   510                 530                 550                 570                 590                 610
TCCTCCAGCAGCAGTGCAGCCCTGTGGCAATGCCACAACGTCTTGCTAGGTCGCAAATGTTGCAGCAGAGCAGTTGCCATGTGATGCAACAACAATGTTGCCAGCAGTTGCCGCAAATCC
 L  Q  Q  Q  C  S  P  V  A  M  P  Q  R  L  A  R  S  Q  M  L  Q  Q  S  S  C  H  V  M  Q  Q  Q  C  C  Q  Q  L  P  Q  I  P

   630                 650                 670                 690                 710                 730
CCCAGCAATCCCGCTATGAGGCAATCCGTGCTATCATCTACTCCATCATCCTGCAAGAACAACAACAGGTTCAGGGTTCCATCCAATCTCAGCAGCAGCAACCCCAACAGTTGGGCCAAT
 Q  Q  S  R  Y  E  A  I  R  A  I  I  Y  S  I  I  L  Q  E  Q  Q  Q  V  Q  G  S  I  Q  S  Q  Q  Q  Q  P  Q  Q  L  G  Q  C

   750                 770                 790                 810                 830                 850
GTGTTTCCCAACCCCAACAGCAGTCGCAGCAGCAACTCGGGCAACAACCTCAACAACAACAATTGGCACAGGGTACCTTTTTGCAGCCACACCAGATAGCTCAGCTTGAGGTGATGACTT
 V  S  Q  P  Q  Q  Q  S  Q  Q  Q  L  G  Q  Q  P  Q  Q  Q  Q  L  A  Q  G  T  F  L  Q  P  H  Q  I  A  Q  L  E  V  M  T  S

   870                 890                 910                 930                 950                 970
CCATTGCGCTCCGTATCCTGCCAACGATGTGCAGTGTTAATGTGCCGTTGTACAGAACCACCACTAGTGTGCCATTCGGCGTTGGCACCGGAGTTGGTGCCTACTGATAAGGAAAGATCT
 I  A  L  R  I  L  P  T  M  C  S  V  N  V  P  L  Y  R  T  T  T  S  V  P  F  G  V  G  T  G  V  G  A  Y  *  *


   990                 1010                1030                1050                1070                1090
CTAGTAATATATAATTGGGTCACCGTTGTTTAGTCGATGGATATGTCGATGCAGCGGTGACAAATAAAGTGTCACACAATGTCATGTGTGACCCGCCCAAACTAGTTGTTTAAATTCTGA
                                                                   I
   1110                1130                1150                1170                1190                1210
AATAAAATAAAAATAAAGTTGTATCAAGACAATGTTCATATTGGCATTGTGTGGATGTCAATCTGATTGCCATGCTTGCAAGTTCATAAGTTTGTCTTTCCTTGTCACAAGCGCAACCTG
 II   III   IV          1  2                                 3
   1230                1250                1270                1290                1310                1330
GTGCCTTAATTAATTATCAATGTACTGGAATAATCACTATTTAAATATAATAGTGTCACTGTAAAATTTGGGTTGAACTCTTTATTGGTTGGAGATTTGAGATCTTGTTTTTTATTGGTT
                                                                                                                  4
   1350                1370                1390                1410                1430                1450
TGTATCTAGTACCAACTAGTACCACTGTCTACAGTGACAATCACAACCAGTTTTGTTGCTAATTATGTTTGCTTTCTTGCAAATACATTCATTGATTCTTTGCATGTCAAATTATTTGCC

   1470                1490                1510                1530                1550                1570
GAGAAAAATACCAGCTTTTTCTTGAACTTCACAAACACCACTAGCTAGTTAATTTGAAATAAGCTGCCAATTTATTTCCTTGTATTTATGCCCTTTACATCGAACCAGTACAATCATTGT

   1590                1610                1630                1650                1670                1690
GCTAGCTAGAGAAAAGGAGTTGCTGGCACGTTCATGTCTAGAACTAAATCAGACTGAATTGAACTTGTTCTTTACAACTAGCTAGGAGCAGAAAACTTACTGTGAATTGCACTTGATTAT

   1710                1730                1750                1770                1790                1810
ATTTGATTTTGTTTAAGAGCTACAGGGAAAAATCATCAAAAAATGCCCTAAAATATCTTATCGTAAAATAGGTTCACCAGGAGATACAGTCGCAAGAAAAAAAAGGCTCACCAGGAGATA

   1830                1850                1870                1890                1910                1930
GATGCTGCACATTATCATTTAGAAATTAGACCAAGAGAAATGTGGATTGGCGGCTCATAATCTATAAGACTATGAACAGGTTGTCCAATCATTTAGAAATTAGATCAAGAAAATGTGGAA

   1950                1970                1990                2010                2030                2050
CTAAACATATTAATGGCTAGGATCATATTTGCACTGTGGGAACTTGGTATGCGCGCTCCAACATGCCCGTTGAAAGCATTTTTTCCTCTGATGCACATATGAAATATGACTGGTCGAAAA

   2070                2090                2110                2130                2150                2170
CAGAAGAAACAGTGAGATAATATGGATGCTTGGATGTTGTAACAAACATAAACTTATTATATATATATCCAATCCACCGAGCGGACGAACATCCTAGGAGCTTTTCTCTCGGAGATCGAA

   2190                2210
AAGAAACCAAATGTCCACGCCAACTGCTTCTTCTCCATCTGATAA
```

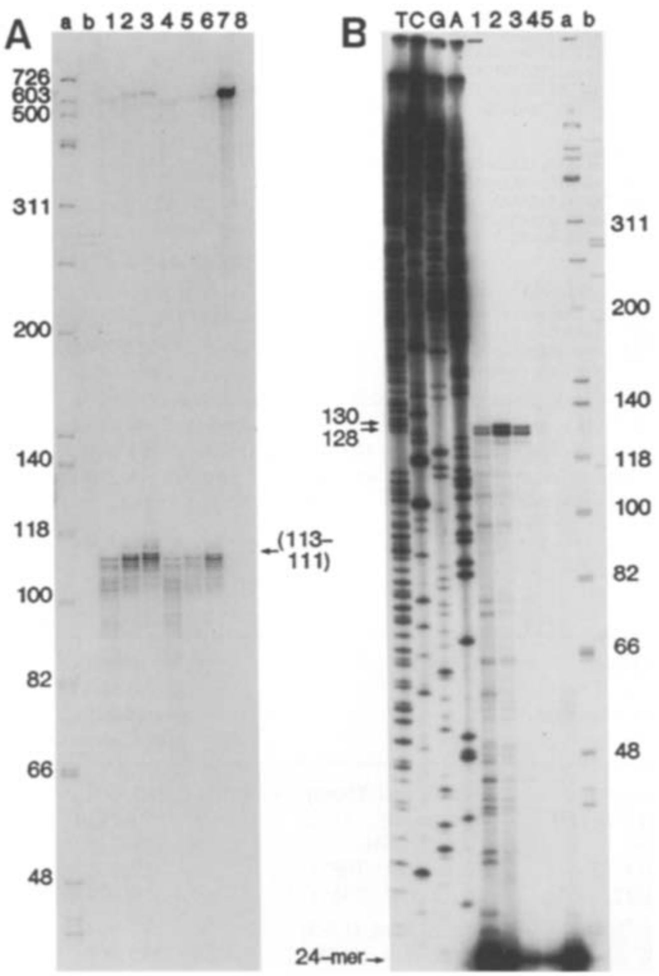


Fig. 3A and B. Determination of the transcription start site of the *LMWG-1D1* gene. **A** S1 nuclease mapping. A ^{32}P-labelled probe of single stranded DNA (position +110 to −498 in Fig. 2) was incubated overnight at 45° C with poly(A)$^+$RNA from wheat endosperm (after brief boiling) and the resultant hybrid was analysed by digestion with S1 nuclease under various conditions. Digestion products were analysed by gel electrophoresis (6% polyacrylamide, 6 M urea). a, b; size markers, in bases; 1–3, S1 nuclease digestion at 30° C; 4–6, S1 nuclease digestion at 37° C; 7, no S1 nuclease; 8, no poly(A)$^+$RNA. S1 nuclease concentrations: 1, 4 units/µl; 2, 4, 2 units/µl; 3, 5, 1 unit/µl; 6, 0.5 unit/µl. **B** Primer extension. Hybridization of the primer to 1 µg of poly(A)$^+$RNA from wheat endosperm was performed at 60° C, after brief boiling. The concentration of primer relative to RNA was calculated assuming that the *LMWG-1D1* transcript represents 1% of wheat endosperm poly(A)$^+$RNA. Primer extension products were analysed by gel electrophoresis (6% polyacrylamide, 6 M urea). T, C, G, A, products of sequencing reactions run in parallel with the primer extension products to allow for precise sizing; 1–4, 50, 10, 5 and 1 × excess of primer, respectively, relative to the estimated *LMWG-1D1* transcript concentration; 5, no RNA; a, b, size markers, in bases

eral direct and inverted repeats. In particular, the v-JUN/GCN4 sequence motif is embedded in an element, also duplicated, that is characteristic of many 5′ flanking regions of endosperm specific genes of wheat, barley and maize (endosperm boxes I and II in Fig. 2; Forde et al. 1985b). As other examples, we have shown in Fig. 2 a set of four direct repeats interspersed with homopolymers of A and regularly spaced every three DNA turns. The first feature of this set of four direct repeats suggest that, at least in vitro, it will be subject to a particular bending behaviour (Koo et al. 1988; Peticolas et al. 1988; Calladine et al. 1988) while the second feature indicates that the four repeats are stereospecifically aligned. Finally, a nearly perfect element of dyad symetry can be identified between the two endosperm boxes, centered at position −324 (Fig. 2).

As in the rest of the *LMWG-1D1* sequence shown in Fig. 2, most of the 5′ flanking region is depleted of CpG dinucleotides, apart from a 185 bp long CpG-rich fragment at its 5′ border (between positions −938 and −753 in Fig. 2). Indeed, the sequence from position −753 to +2226 contains, with reference to its base composition, only about half of the expected number of CpGs (58 vs 122) and about 1.4 times as much the expected number of TpGs (202 vs 140). However, the 185 bp 5′ border fragment alone contains 19 CpGs which roughly equals the number of GpCs (16), and hence does not show any depletion. The significance of these observations in relation to gene activity will be presented briefly in the discussion.

3′ flanking region. Four AATAAA polyadenylation signals are found within the 1.2 kb 3′ flanking sequence. The first motif is located 80 nucleotides 3′ of the 2 stop codons, while the last 3 motifs are clustered another 60 nucleotides further downstream (Fig. 2). Comparison of the 3′ termini of 4 cDNAs similar to the *LMWG-1D1* sequence revealed several different polyadenylation sites 3′ of the clustered AATAAA motifs, but none between these and the first motif (Okita et al. 1985 and Fig. 2). In particular, cDNA B11–33 extends well beyond the clustered polyadenylation signals (polyadenylation site Number 4 in Fig. 2). A comparison of consensus 3′ flanking sequences corresponding to the four prolamin gene families that we analyse in the next section reveals a strong conservation of sequence around two AATAAA motifs but not elsewhere, and thus suggests that these two motifs are maintained by selection for function (Fig. 4).

Comparison of LMWG-1D1 to other prolamin genes

The size and the amino acid composition of the predicted LMWG-1D1 protein indicates that this sequence corresponds to either an α,β-gliadin, a γ-gliadin, or a LMW glutenin but not to an ω-gliadin or a HMW glutenin (data not shown). The location of the *LMWG-1D1* DNA sequence and of sequences related to pTag544 on the group 1 homeo-

Fig. 2. Nucleotide sequence of 3.2 kb of the 3.5 kb *Bam*H1-*Sst*1 fragment of clone LMWG-1D1. The nucleotide sequence is shown aligned with the predicted protein sequence. Numbering is from the transcription start site, at position 0 (see Fig. 3). Some potential eukaryotic *cis*-acting signals of gene expressions are *boxed.* Note that for the 2 endosperm boxes, only the v-JUN/GCN4 binding sites present in the 3′ part are known regulatory signals. The sites of polyadenylation of 4 cDNA clones related to LMWG-1D1 are shown: 1, pTag544; 2, pB312; 3, pB48; 4, pB11–33 (see Table 1 for references). Several repeats of the 5′ flanking sequence are *overlined.* The derived protein sequence has been divided into a signal peptide, closely followed by a repeat domain rich in proline and glutamine residues and a C-domain which is mostly non-repetitive

```
              1                                                                                     87
   LMW GLU.  ...TGAtAAggaaag.tctctagtAaT..atata.GttggatCAcCGTT..g.TttAgtc.gATGgatatgTCGaTGcAgCGgTGAc
    B HORD.  taaTGAtAAgaaaaggtctctagaAaT..atata.Gttg-atCAcCGTT..g.TctAatc.gATGtatatgTCGaTGtAgCGgTGAc
 GAMMA GLI.  ...TGAaAA..acgcaagagc-atAcTaataggtaGatggatCAtCGTT..gcTt.Agct.gATG-accaaTCGaTGtAaCGaTGAc
 ALPHA GLI.  ...TGAgAAgagaagaactctagtAcTagatatatGaaa...CAcCGTTtt.cTt.Agtcc.ATGgtttggTCGtTGtAgCGgTGAa
             ---TGA-AA---------------A-T--------G------CA-CGTT----T--A-----ATG------TCG-TG-A-CG-TGA-

            88                                                                                     174
   LMW GLU.  A.AATAAAGTGtCacaCAacgTcATGTgtGAcCc..gcCcaaagtaCTAGTTgtttAAattcTGa.AATAAAtAcAAAtAaAgtTg
    B HORD.  A.AATAAAGTGtCacaCAaccTtATGTgtGAcC..ggcCcaaa...CTAGTTgtttAAattcTGa.AATAAAtAtAAAtAaAgtTc
 GAMMA GLI.  A.AATAAAGTGgCgtgCAccaTcATGTgtGA-Cc-gacC..aagtgCTAGTT...cAAgactTGgAATAAAAgAcAAAcA-AgtTc
 ALPHA GLI.  AaAATAAAGTGaCatgCActaTcATGTaaGAaCccgaaCt-..ataCTAGTT...cAAa.ctTGggAATAAAAgAcAAAcAcAtgTc
             A-AATAAAGTG-C---CA---T-ATGT--GA-C-----C-------CTAGTT----AA----TG--AATAAA-A-AAA-A-A--T-

           175                                                                                     261
   LMW GLU.  tactc-agacaatgttcatattggcattgtgtggatgtc-atctga-t-ccatgcttgcaagttcataagtt-gtctt-cc--gtca
    B HORD.  atgatgacta-ctg-aaagtttctc-aacaagt-gaa---tgtattaattc------------ccaaac-gaa-gacta--tgaaa
 GAMMA GLI.  -tgtttgcca-catt-cttgtcattg.ttccattcactg.tgtatttagat--gttcatccctaactacaattctag-cttacacat
 ALPHA GLI.  ttgtctacatatgattgtttgtttga-ttccattcat.g.tgc-c-t-cacaagttcacccctaatt-tatatattatgcattaagt
             ---------------------------------------------------------------------------------------
```

Fig. 4. Comparison of the consensus 3′ flanking sequences of 4 prolamin gene families (see Table 1 and Fig. 5 for details of the 4 families). LMW GLU, LMW glutenin; B HORD, B hordein; GAMMA GLI, γ-gliadin; ALPHA GLI, α,β-gliadin. *Dots* correspond to gaps introduced to optimise the alignment. *Dashes* indicate lack of consensus. The bottom line of the alignment shows only nucleotides common to the 4 families. *Overlined* are the 2 AATAAA sequence motifs that are conserved among the 4 families. A patchwork of limited sequence similarity (i.e. between 2 or 3 families only) can be observed outside the 2 highly conserved regions centered on the AATAAA motifs

Table 1. C-domain of prolamin genes of wheat and barley

Putative encoded protein type	EMBO library code of the gene sequence	C: cDNA G: genomic	Original name of the gene sequence	Reference	Position of first and last nucleotides of C-domain[a]
Wheat	WHTGLG	C	pTag544	Bartels and Thompson (1983)	103–663
LMW glutenin	–	G	*LMWG-1D1*	This paper	1350–1910
	WHTGLIGBC[b]	C	pB48	Okita (1984)	178–738
	WHTGLIGBA[b]	C	pB11–33	Okita et al. (1985)	367–966
	WHTGLIGBB[b]	C	pB312	Okita et al. (1985)	377–976
Barley	–	G	λhor2-4	Brandt et al. (1985)	830–1372
B hordein	BLYB3HORD	C	pB7	Forde et al. (1985a)	262–801
	BLYHORB	G	pBHR184	Forde et al. (1985b)	912–1451
	BLYB1HORD	C	pB11	Forde et al. (1985a)	200–739
Wheat	WHTGLGAP	G	L311A	Rafalski (1986)	990–1502
gamma-gliadin	WHTGLGB	G	L311B	Rafalski (1986)	897–1382
	WHTGLIGY	G	pW1621	Sugiyama et al. (1986)	622–1110
	–	C	pTag1436	Bartels et al. (1986)	493–969
Wheat	WHTGLIABD	G	pW8142	Sumner-Smith et al. (1985)	1273–1866
alpha, beta-gliadin	WHTGLIABC	C	pA42	Okita et al. (1985)	394–996
	WHTGLIABG	C	pA735	Okita et al. (1985)	436–966
	WHTGLIABB	G	pW1215	Sumner-Smith et al. (1985)	954–1487
	WHTGLIABF	C	pA26	Okita et al. (1985)	394–840
	WHTGLIABH	C	pA1235	Okita et al. (1985)	399–905
	WHTGLIABA	C	pB212	Okita et al. (1985)	409–942

[a] Including 3 or 6 nucleotides 3′ past the last codon (see Fig. 5)

[b] These sequences were originally classified as γ-gliadin type sequences by Okita et al. (1985)

logous chromosomes (Harberd et al. 1985 and Fig. 1B) rules out that these sequences encode, α,β-gliadins, whose genes are on the chromosomes of group 6. As a result, our sequence and sequences related to pTag544 most probably correspond to genes at the *gli-1* locus and encode either γ-gliadins, or LMW glutenins (Payne et al. 1984).

A sequence similar to the 15 amino acids long amino-terminal consensus of LMW glutenin subunits is found towards the amino-terminus of the predicted LMWG-1D1 protein (Shewry and Miflin 1984). Since this sequence bears little resemblance to the consensus amino-terminus of γ-gliadins (Kasarda et al. 1984), we believe that the coding region of *LMWG-1D1* specifies a LMW glutenin subunit.

A dot matrix comparison (program DOTPLOT, data not shown) of DNA sequences encoding various prolamin types revealed that our sequence is closely related to the cDNA γ-gliadin type sequences of Okita et al. (1985) and to barley B1 and B3 hordein sequences but less to γ-gliadin and α,β-gliadin sequences.

To analyse further the extent of similarity between *LMWG-1D1* and other prolamin gene sequences we have compared in detail the C-domain which follows the repeat domain of 20 prolamin genes corresponding to the types mentioned above. We reasoned that the sequence, mostly non-repetitive in nature, that is found 3′ of the repeat domain of these genes will be less susceptible to ambiguous

Table 2. Comparison of C-terminal domains of prolamin genes

Sequence type of C-domain	Wheat LMW glutenin	Barley B hordein	Wheat gamma-gliadin	Wheat alpha, beta-gliadin
Wheat	86.3%	79.6%	59.0%	52.4%
LMW glutenin	99.8%[a]	84.0%	63.9%	61.0%
Barley		88.5%	58.3%	52.7%
B hordein		99.6%[a]	63.9%	57.3%
Wheat			88.5%	58.6%
gamma-gliadin			95.6%	63.7%
Wheat				87.8%
alpha, beta-gliadin				96.8%

Lowest and highest scores of identity within and between families

[a] Despite near identity of certain C-domain sequences within a family, the corresponding genes differ in their repeat domain, by the gain or loss of repeat units and/or simple sequences

alignments than the repeat domain and could therefore provide a clearer classification of prolamin gene families. Also the coding region 5′ of the repeat domain of the prolamin sequences compared here is too small to be useful for this purpose.

After identification of the C-domain for each of the 20 different prolamin genes (see Table 1), optimal pairwise alignments of the delineated sequences were obtained using the computer program GAP (gap weight = 5.0, gap length = 0.1). Coefficients of identity were defined as the percentage of identical nucleotides over the number of aligned positions, and gaps were ignored in that calculation. On this basis we distinguished 4 families of sequences within which each pairwise comparison gave a score over 85% (data summarized in Table 2). As expected, all sequences of genes identified as α,β-gliadins fell into one family. However, the sequences designated as γ-gliadin type by Okita et al. (1985) were found together with *LMWG-1D1* to be in a family that is distinct from that containing the γ-gliadin cDNA pTag1436 (Bartels et al. 1986) but is closely related to B hordein sequences.

For each family, the pairwise alignments were then combined, with slight modifications (programs GAP and PRETTY). Finally, the 20 DNA sequences of the 4 families were optimally aligned. This was done under the two conservative constraints of maintaining the pairwise alignments of each family and minimizing the number of nucleotide differences between families (programs GAP, LINEUP and PRETTY combined with manual alignment). We have shown in Fig. 5 an amino acid translation of the overall alignment which emphasizes both the similarity and the divergence of gliadins, LMW glutenins and B hordeins. In particular, it distinguishes two regions of similarity between the four families (subdomains I and II), as well as several regions which are variable within and between families. Inspection at the DNA level (not shown) and at the amino acid level of the LMW glutenin and the α,β- and γ-gliadin families reveals many differences between them, both in the similar and the variable regions of the C-domain. Thus, LMW glutenin and γ-gliadin genes, although found at the same location are not more closely related to each other than to the α,β-gliadin genes found on different chromosomes. Our global alignment also reveals an extensive size polymorphism within three of the four families (LMW glutenin, α,β- and γ-gliadin), the length variants being observed mostly in two regions which are glutamine-rich and/or show discernible repeat patterns. Thus, length variants within the C-domain result predominantly from the gain or loss of simple sequences or repeat motifs.

A similar gain or loss of simple sequences or repeat motifs is observed within the repeat domain that precedes the C-domain. This is exemplified by a comparison of the *LMWG-1D1* sequence and of cDNA B11–33 (Fig. 6). Starting 33 nucleotides 3′ of the Arg codon that corresponds to the putative mature N-terminal end, the repeat domain of *LMWG-1D1* contains 13 imperfect repeat units based on the heptamer motif CAG CAA CAA CMA CCA TTT YCA, where M = C or A and Y = C or T. Individual repeats vary in length, such that alignment of the 13 repeats reveals the gain or loss of codons in the first half of the repeat motif. Also, the repeat unit appears at the 3′ end of the domain closer to a 14-mer made up of two variants of the heptamer motif. Comparison of the *LMWG-1D1* repeat domain with that of *B11–33* identifies the gain or loss of one CAA codon within a $(CAA)_n$ simple sequence stretch and of two whole heptamers. Finally, the repeat motifs are specific both in length and in sequence to each of the four families described in Fig. 5 (data not shown).

Discussion

Clones were isolated from a wheat genomic DNA library that belong to a small prolamin gene family located on the group 1 chromosomes of wheat (Fig. 1). Sequence analysis of clone LMWG-1D1 showed that, like other prolamin genes (Kreis et al. 1985a), its coding region possesses a proline and glutamine rich domain encoded by a tandem array of repeats followed by a unique sequence domain interspersed with several stretches of glutamine codons (Fig. 2). Several lines of evidence suggested that the *LMWG-1D1* gene encodes a LMW glutenin subunit.

The classification of *LMWG-1D1* in relation to other prolamin gene sequences was approached primarily through an analysis of the unique sequence (C-domain) which follows that encoding the proline and glutamine rich domain of prolamins. Four families were distinguished from a comparison of 20 different prolamin sequences (Fig. 5). *LMWG-1D1* and other sequences were found to be members of a family, now called the LMW glutenin gene family, that is distinct from the γ-gliadin and α,β-gliadin gene families. Furthermore, the LMW glutenin gene family shows extensive similarity with the barley B hordein family. Comparisons of the repeat domain and of the non-coding 5′ flanking sequence of *LMWG-1D1* with that of other prolamin genes also lead to the distinction of the same 4 families (data not shown). Hybridization data from several sources (Harberd et al. 1985; Bartels et al. 1986; Forde et al. 1985a) indicate that these 4 gene families contain between 2 and 20 copies per haploid component of each diploid genome.

The similarity between the four families suggests either that they have a common origin and have resulted from the divergence of four ancestral sequences that were then amplified more recently; or that divergence occurred within an ancient family containing many members that have since evolved into sub-families by homogenisation of different copies through DNA turnover mechanisms (Dover and Tautz 1986). Such mechanisms may also account for the

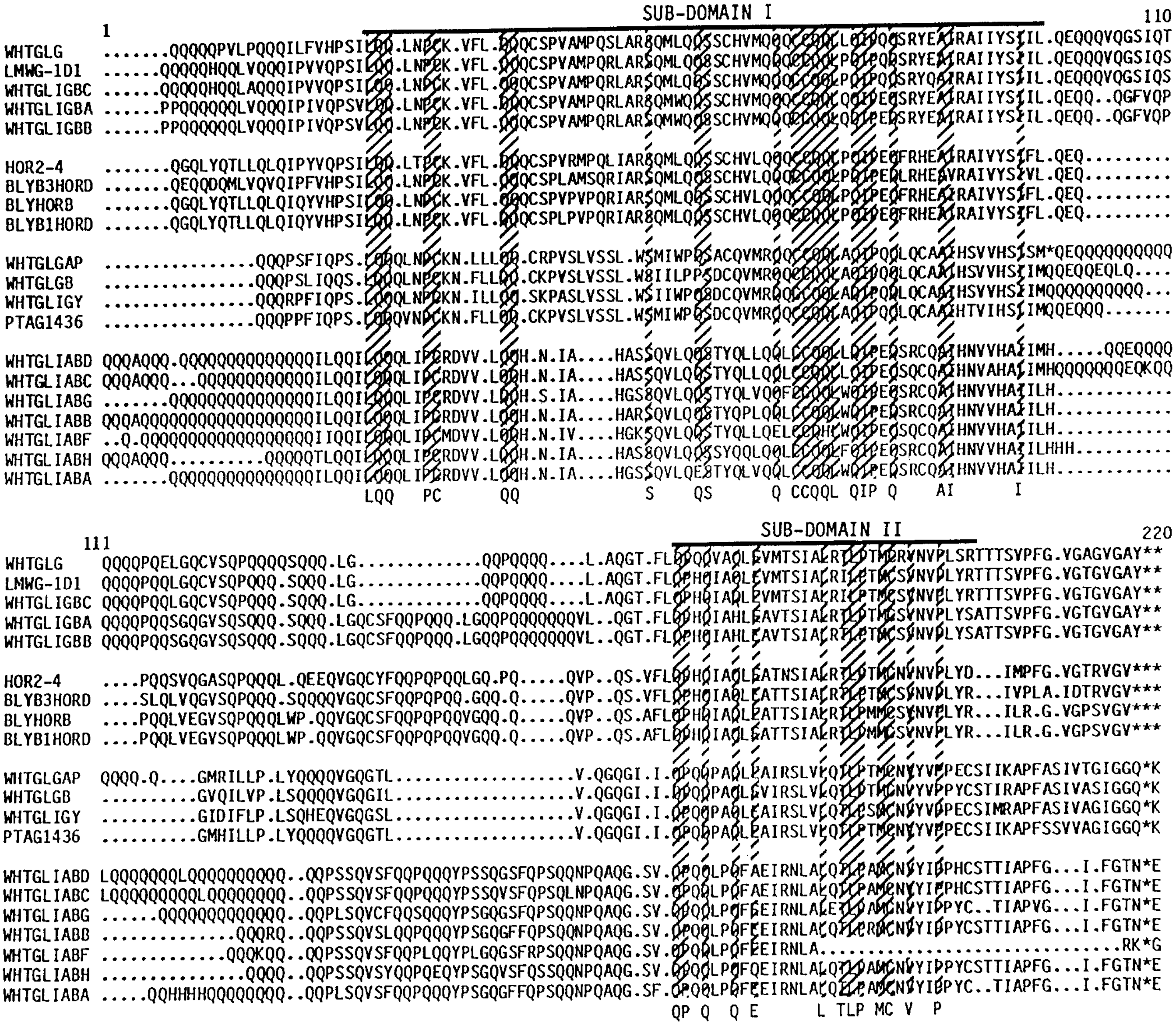


Fig. 5. Amino acid comparison of derived protein sequences of the C-domain of several prolamin genes of wheat and barley. The sequences under comparison cover roughly 200 amino acids and are defined and listed in Table 1. Their grouping in 4 families has been obtained from multiple pairwise alignments, as summarized in Table 2. The 4 families are, from top to bottom: LMW glutenin (wheat), B hordein (barley), γ-gliadin (wheat) and β-gliadin (wheat). Two regions, subdomain I and II (*overlined*), of clear similarity between the 4 families were recognized in the alignment. Conserved amino acids in these 2 regions are indicated by *hatches* and are noted below the alignment. *Dots* correspond to gaps introduced to optimize the alignment. *Stars* indicate stop codons

maintenance of homogeneity among the amplified sequences of the former model, if amplification were an old event. However, irrespective of the model chosen, our comparison indicates that the divergence between LMW glutenin and B hordein sequences is of more recent origin than the divergence between sequences of any other pair from the four families. Since cytological and genetical data (reviewed in Kreis et al. 1985b) provides further evidence that LMW glutenin and B hordein genes are orthologous, their divergence may have begun simultaneously with the separation of the ancestral diploid wheat and barley species.

The alignment and grouping of the prolamin amino acid and nucleotide sequences (Fig. 5) clearly illustrate that besides single base changes, many mutations resulting from deletion/addition of many bases or repeat units and base reiteration, probably by slippage replication, have accumulated within these gene families. Indeed these latter types of mutations greatly facilitated the separation of the 20 prolamin sequences into 4 families. The consequences of unequal cross-over and slippage replication events appear to be tolerated extensively in some regions, (the repeat domain and the variable region of the C-domain) but less frequently in other regions (the N-terminal end and the subdomains I and II of the C-domain) and to be the principal sources of polypeptide length polymorphisms within and between the prolamin families. This may imply that there are certain constraints imposed on the evolution of specific regions of storage proteins. In particular, the sequence CCQQL found in subdomain I is also conserved among several other plant storage proteins as well as among

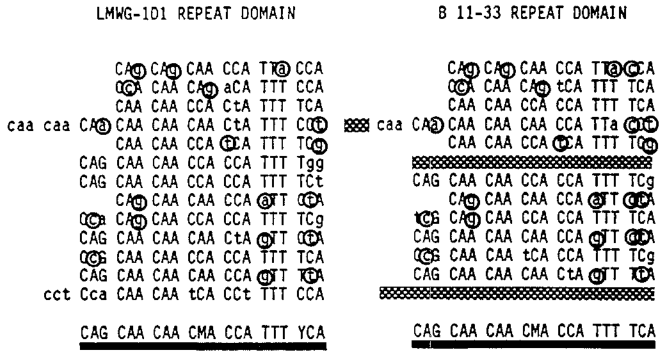

LMWG-1D1 REPEAT DOMAIN
CAg CAg CAA CCA TTa CCA
cCA CAA CAg aCA TTT CCA
CAA CAA CCA CtA TTT TCA
caa caa CAa CAA CAA CAA CtA TTT Cct
CAA CAA CCA tCA TTT TCg
CAG CAA CAA CCA CCA TTT Tgg
CAG CAA CAA CCA CCA TTT TCt
CAg CAA CAA CCA aTT ctA
cCa CAg CAA CCA CCA TTT TCg
CAG CAA CAA CAA CtA gTT ctA
cCG CAA CAA CCA CCA TTT TCA
CAG CAA CAA CAA CCA gTT TtA
cct Cca CAA CAA tCA CCt TTT CCA
CAG CAA CAA CMA CCA TTT YCA
B 11-33 REPEAT DOMAIN
CAg CAg CAA CCA TTa cCA
cCA CAA CAg tCA TTT TCA
CAA CAA CCA CCA TTT TCA
caa CAa CAA CAA CAA CCA TTa cct
CAA CAA CCA tCA TTT TCg
CAG CAA CAA CCA CCA TTT TCg
CAg CAA CAA CCA aTT cCA
tCG CAg CAA CCA CCA TTT TCA
CAG CAA CAA CAA CCA gTT cCA
cCG CAA CAA tCA CCA TTT TCg
CAG CAA CAA CAA CtA gTT TtA
CAG CAA CAA CMA CCA TTT TCA

Fig. 6. Repeat structure of the two DNA sequences encoding the proline and glutamine rich domains of the predicted LMWG-1D1 and B11–33 (Okita et al. 1985) proteins. Both sequences are arranged to reveal their repetitive nature. The consensus repeat unit of each sequence is shown *underlined* below each array. Nucleotides differing from the consensus are shown in *lower case* and are *circled* when shared by the *LMWG-1D1* and *B11–33* sequences. The *shaded areas* in the *B11–33* sequence correspond to extra nucleotides in the *LMWG-1D1* sequence

many plant protease inhibitors. Therefore, we can speculate that this sequence prevents premature degradation of storage proteins.

The major seed storage proteins of cereals are accumulated exclusively in the endosperm tissue. Studies of steady state mRNA levels have indicated that a coordinate endosperm-specific expression of prolamin genes corresponds to this synthesis (Rahman et al. 1984; Bartels and Thompson 1986; Reeves et al. 1986). Recently, we have shown that the 1 kb 5′ flanking sequence of *LMWG-1D1* directs endosperm specific expression of a reporter gene (*cat*) in seeds of transgenic tobacco plants. Furthermore, a deletion analysis of that sequence indicated that a DNA fragment located between positions −326 and −160 in Fig. 2 is necessary to confer that activity (Colot et al. 1987). Here we have shown by S1 nuclease mapping and primer extension that *LMWG-1D1* is almost certainly active in wheat.

Several sequence motifs are present in the 5′ flanking region of gene *LMWG-1D1* that have been associated with gene expression in other eukaryotic systems. They included a TATA box, a CCAT box, several SV40 enhancer sequence motifs and two v-JUN/GCN4 binding sites. Recent studies have shown that many of these *cis*-acting motifs bind interchangeable mammalian and yeast *trans*-acting factors of gene expression (Buratowski et al. 1988; Cavallini et al. 1988; Guarente 1988; Struhl 1987; 1988), which suggests that transcriptional activation shares common mechanisms in many eukaryotes. Also, there is growing evidence that the specifity of gene expression is, at least in part, determined by the combinatorial arrangement of binding sites and/or *trans*-acting factors (Pfeifer et al. 1987; Schirm et al. 1987; Cato et al. 1988; Ondek et al. 1988). Thus, we are currently investigating whether the v-JUN/GCN4 binding site and other sequences present in the 160 bp fragment of *LMWG-1D1*, which is sufficient to confer endosperm-specific expression of a chimaeric gene in transgenic tobacco (Colot et al. 1987; V. Colot, unpublished observations), bind protein factors in wheat and tobacco.

Several other structural features of the *LMWG-1D1* gene have been described which may be relevant to its expression. In particular, the depletion of CpG dinucleotides along most of the characterised DNA except for its 5′ end may reflect the requirement for as few methylated cytosine sites as possible in order that the gene is expressed (Cedar 1988; Antequera and Bird 1988). Alternatively, CpG depletion may simply be a consequence of stochastic processes over long time periods in which methylated cytosines, subject to spontaneous deamination, have been replaced by thymidines.

The comparison of consensus 3′ flanking sequences revealed 2 highly conserved AATAAA motifs, within the first 160 nucleotides downstream of the stop codon (Fig. 4). However, a comparison of our sequence with that of related cDNAs suggests that polyadenylation occurs at different sites (Fig. 2). Similar observations have been made with other plant genes (Dean et al. 1986). Thus, sequences other than the polyadenylation signal AATAAA are necessary to specify the endonucleolytic cleavage of the LMW glutenin precursor RNAs.

Acknowledgements. We would like to thank Dr. G. Murphy for introducing us to the different computer programs. V.C. was the recipient of an INRA (France) studentship and is currently supported by a fellowship from Sidney Sussex College, Cambridge. This work was also carried out in the framework of BAP contract 0106 UK from the Commission of the European Communities.

References

Antequera F, Bird AP (1988) Unmethylated CpG islands associated with genes in higher plant DNA. EMBO J 7:2295–2299

Bartels D, Thompson RD (1983) The characterization of cDNA clones coding for wheat storage proteins. Nucleic Acids Res 11:2961–2977

Bartels D, Thompson RD (1986) Synthesis of mRNAs coding for abundant endosperm proteins during wheat grain development. Plant Sci 46:117–125

Bartels D, Altosaar I, Harberd NP, Barker RD, Thompson RD (1986) Molecular analysis of gamma-gliadin gene families at the complex *Gli-1* locus of bread wheat (*T. aestivum* L.). Theor Appl Genet 72:845–853

Brandt A, Montenbault A, Cameron-Mills V, Rasmussen SK (1985) Primary structure of a B1 hordein gene from barley. Carlsberg Res Commun 50:333–345

Buratowski S, Hahn S, Sharp PA, Guarente L (1988) Function of a yeast TATA element-binding protein in a mammalian transcription system. Nature 334:37–42

Calladine CR, Drew HR, McCall MJ (1988) The intrinsic curvature of DNA in solution. J Mol Biol 201:127–137

Cameron JR, Philippsen P, Davies RW (1977) Analysis of chromosomal integration and deletions of yeast plasmids. Nucleic Acids Res 4:1429–1448

Cato ACB, Skroch P, Weinmann J, Butkeraitis P, Ponta H (1988) DNA sequences outside the receptor-binding sites differentially modulates the responsiveness of the mouse mammary tumour virus promoter to various steroid hormones. EMBO J 7:1403–1410

Cavallini B, Huet J, Plassat J-L, Sentenac A, Egly JM, Chambon P (1988) A yeast activity can substitute for the HeLa cell TATA box factor. Nature 334:77–80

Cedar H (1988) DNA methylation and gene activity. Cell 53:3–4

Colot V, Roberts LS, Kavanagh TA, Bevan MW, Thompson RD (1987) Localization of sequences in wheat endosperm protein genes which confer tissue-specific expression in tobacco. EMBO J 6:3559–3564

Dean C, Tamaki S, Dunsmuir P, Faveau M, Katayama C, Dooner H, Bedbrook J (1986) mRNA transcripts of several plant genes are polyadenylated at multiple sites *in vivo*. Nucleic Acids Res 14:2229–2240

Devereux J, Haeberli P, Smithies O (1984) A comprehensive set of sequence analysis programs for the VAX. Nucleic Acids Res 12:387–395

Dover GA, Tautz D (1986) Conservation and divergence in multigene families: alternatives to selection and drift. Philos Trans R Soc Lond [Biol] 312:275–289

Entwistle J (1988) Primary structure of a C-hordein gene from barley. Carlsberg Res Commun 53:247–258

Fedoroff N (1983) Notes on cloning maize DNA. Maize Genet Co-op Newslett 57:154 (Add)

Forde BG, Kreis M, Williamson MS, Fry RP, Pywell J, Shewry PR, Bunce N, Miflin BJ (1985a) Short tandem repeats shared by B- and C-hordein cDNAs suggest a common evolutionary origin for two groups of cereal storage protein genes. EMBO J 4:9–15

Forde BG, Heyworth A, Pywell J, Kreis M (1985b) Nucleotide sequence of a B1 hordein gene and the identification of possible upstream regulatory elements in endosperm storage protein genes from barley, wheat and maize. Nucleic Acids Res 13:7327–7339

Guarente L (1988) UASs and enhancers: common mechanism of transcriptional activation in yeast and mammals. Cell 52:303–305

Halford NG, Forde J, Anderson OD, Greene FC, Shewry PR (1987) The nucleotide and deduced amino acid sequences of a HMW glutenin subunit gene from chromosome 1B of bread wheat (*Triticum aestivum* L.) and comparison with those of genes from chromosomes 1A and 1D. Theor Appl Genet 75:117–126

Harberd NP, Bartels D, Thompson RD (1985) Analysis of the gliadin multigene loci in bread wheat using nullisomic-tetrasomic lines. Mol Gen Genet 198:234–242

Henikoff S (1984) Unidirectional digestion with exonuclease III creates targeted breakpoints for DNA sequencing. Gene 28:351–359

Kasarda DD, Lafiandr D, Morris R, Shewry PR (1984) Genetic relationships of wheat gliadin proteins. Kulturpflanze 32:41–60 (Suppl)

Koo H-S, Crothers DM (1988) Calibration of DNA curvature and a unified description of sequence-directed bending. Proc Natl Acad Sci USA 85:1763–1767

Kovacs BJ, Butterworth PHW (1986) The effect of changing the distance between the TATA-box and Cap site by up to three base pairs on the selection of the transcriptional start site of a cloned eukaryotic gene *in vitro* and *in vivo*. Nucleic Acids Res 14:2429–2441

Kozak M (1986) Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes. Cell 44:283–292

Kreis M, Forde BG, Rhaman S, Miflin BJ, Shewry PR (1985a) Molecular evolution of the seed storage proteins of barley, rye and wheat. J Mol Biol 183:499–502

Kreis M, Shewry PR, Forde BG, Forde J, Miflin BJ (1985b) Structure and evolution of seed storage proteins and their genes with particular reference to those of wheat, barley and rye. Oxford Surveys Plant Mol Cell Biol 2:253–317

Loenen WAM, Blattner FR (1983) Lambda Charon vectors (Ch 32, 33, 34 and 35) adapted for DNA cloning in recombination deficient hosts. Gene 26:171–179

Maniatis T, Fritsch EF, Sambrook J (1982) Molecular cloning: a laboratory manual. Cold Spring Harbour Laboratory Press, Cold Spring Harbor, New York

Messing J (1983) New M13 vectors for cloning. Methods Enzymol 101:20–78

Nasmyth KA (1983) Molecular analysis of a cell lineage. Nature 302:670–676

Okita TW (1984) Identification and DNA sequence analysis of a gamma-type gliadin cDNA plasmid from winter wheat. Plant Mol Biol 3:325–332

Note added in proof

The sequence of gene *LMWG-1D1* will appear in the EMBL/GenBank/DDBJ Nucleotide Sequence Databases under the accession number X13306.

Okita TW, Cheesebrough V, Reeves CD (1985) Evolution and heterogeneity of the alpha-/beta-type and gamma-type gliadin DNA sequences. J Biol Chem 260:8203–8213

Ondek B, Gloss L, Herr W (1988) The SV40 enhancer contains two distinct levels of organisation. Nature 333:40–45

Payne PI (1987) Genetics of wheat storage proteins and the effect of allelic variation on bread-making quality. Annu Rev Plant Physiol 38:141–153

Payne PI, Holt LM, Jackson EA, Law CN (1984) Wheat storage proteins: their genetics and their potential for manipulation by plant breeding. Philos Trans R Soc Lond [Biol] 304:359–371

Peticolas WL, Wang Y, Thomas GA (1988) Some rules for predicting the base-sequence dependence of DNA conformation. Proc Natl Acad Sci USA 85:2579–2583

Pfeifer K, Arcangioll B, Guarente L (1987) Yeast HAP1 activator competes with the factor RC2 for binding to the upstream activation site UAS1 of the CYC1 gene. Cell 49:9–18

Rafalski JA (1986) Structure of wheat gamma-gliadin genes. Gene 43:221–229

Rahman S, Kreis M, Forde BG, Shewry PR, Miflin BJ (1984) Hordein-gene expression during development of a barley (*Hordeum vulgare*) endosperm. Bioche J 223:315–322

Reeves CD, Krishnan HB, Okita TWJ (1986) Gene expression in developing wheat endosperm. Accumulation of gliadin and ADPglucose pyrophorylase messenger RNAs. Plant Physiol 82:34–40

Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA 74:5463–5476

Schirm S, Jiricny J, Schaffner W (1987) The SV40 enhancer can be dissected into multiple segments, each with a different cell type specificity. Genes Devel 1:65–74

Sears ER (1966) Nullisomic-tetrasomic combination in hexaploid wheat. In: Riley R, Lewis KR (eds) Chromosome manipulation and plant genetics. Oliver and Boyd, Edinburgh, pp 29–45

Shewry PR, Miflin BJ, Lew EJ-L, Kasarda DD (1984) The preparation and characterization of an aggregated gliadin fraction from wheat. J Exp Bot 34:1403–1410

Staden R (1982) An interactive graphics program for comparing and aligning nucleic acid and amino acid sequences. Nucleic Acids Res 10:2951–2961

Struhl K (1987) The DNA-binding domains of the jun oncoprotein and the yeast GCN4 transcriptional activator protein are functionally homologous. Cell 50:841–846

Struhl K (1988) The JUN oncoprotein, a vertebrate transcription factor, activates transcription in yeast. Nature 332:649–650

Sugiyama T, Rafalski A, Soll D (1986) The nucleotide sequence of a wheat gamma-gliadin genomic clone. Plant Sci 44:205–209

Sumner-Smith A, Rafalski JA, Sugiyama T, Stoll M, Soll D (1985) Conservation and variability of wheat alpha/beta-gliadin genes. Nucleic Acids Res 13:3905–3916

Thompson RD, Bartels D, Harberd NP, Flavell RB (1983) Characterization of the multigene family coding for HMW glutein subunits in wheat using cDNA clones. Theor Appl Genet 67:87–96

Thompson RD, Bartels D, Harberd NP (1985) Nucleotide sequence of a gene from chromosome 1D of wheat encoding a HMW-subunit. Nucleic Acids Res 13:6833–6846

von Heijne G (1985) Signal sequences: the limits of variation. J Mol Biol 184:99–105

Communicated by H. Saedler

Received August 26, 1988