

Analysis of Seed Storage Protein Genes of Oats*

(Received for publication, February 27, 1990)

Mark A. Shotwell‡, Scott K. Boyer, Ruth S. Chesnut§, and Brian A. Larkins¶

From the Department of Plant Sciences, University of Arizona, Tucson, Arizona 85721

We have isolated genomic clones encoding the two major classes of seed storage proteins in oats, the 12 S globulins and the avenins. The globulin genes encode glutamine-rich, sulfur-poor storage proteins that are highly conserved in sequence and structure. The globulin genes contain three short introns whose positions in the coding sequence are the same as in storage globulin genes in legumes and other dicots. The avenin genomic clone contains four tightly linked genes that belong to both of the two avenin gene subfamilies. The avenin genes encode glutamine-rich, lysine-poor proteins that vary in length due to differences in the number of peptide repeats. Although globulin and avenin genes are expressed coordinately during oat seed development, their promoter regions do not contain any conserved sequence elements that might determine developmental timing. Previous studies showed that there are roughly equal amounts of globulin and avenin mRNAs in developing oat seed, despite there being much more globulin than avenin in mature seed. Storage protein synthesis in oats must therefore be controlled partially by post-transcriptional mechanisms. Sequence analysis of globulin and avenin genes has provided several clues as to why globulin mRNAs may be translated more efficiently than avenin mRNAs.

Cereal seeds contain large amounts of storage proteins that are degraded during germination to provide nutrients to the seedling. Most of these seed storage proteins fall into two classes, the alcohol-soluble prolamines and the saline-soluble globulins (1). The seeds of most cereals contain relatively large amounts of prolamine storage proteins and small amounts of globulin storage proteins. Oat seeds are unusual in this regard in that they store, by weight, only 10–20% prolamines, called avenins, and 75–80% globulins (2, 3).

The oat globulin shares structural features with the 11 S globulins of legumes and other dicots, although it is much less soluble (4). The protein is isolated as a 340-kDa hexamer with a sedimentation coefficient of about 12 S (5). Each of the six subunits consists of a 35-kDa acidic polypeptide disulfide

bonded to a 22-kDa basic polypeptide. The globulin subunits are synthesized on rough endoplasmic reticulum membranes and are proteolytically processed into acidic and basic polypeptides before deposition in the vacuole (6).

The avenins of oats comprise a group of alcohol-soluble proteins ranging in size from 17 to 23 kDa (7). Amino-terminal sequence analysis revealed homology both between avenin polypeptides as well as between the avenins and the prolamines of wheat, rye, and barley (8). More recently, sequences have been determined for several of the polypeptides (9, 10), although many of them are blocked to Edman degradation.

An important consequence of the high-globulin, low-prolamine composition of oat seeds is that oat protein provides a better balance of the amino acids essential for humans and other monogastric animals than does seed protein from most other cereals. We are attempting to gain an understanding of the genetic basis for the high-globulin, low-prolamine composition of oat seeds. We have isolated cDNA clones corresponding to globulin and avenin mRNA sequences (11, 12). Despite the predominance of globulin in the mature seed, we found that nearly equal amounts of globulin and avenin messenger RNA are present in developing oat endosperm (12). The control of storage protein synthesis in oat seeds is clearly complex, involving both transcriptional and post-transcriptional mechanisms. To gain further insight into the regulation of oat seed storage protein synthesis, we have isolated and characterized genomic sequences for both the globulins and avenins of oats.

MATERIALS AND METHODS

Materials—Restriction endonucleases and mung bean nuclease were purchased from Bethesda Research Laboratories, Gaithersburg, MD. [α - ^{32} P]dATP (3000 Ci/mmol) and [α - ^{35} S]dATP (1000 Ci/mmol) were obtained from Du Pont-New England Nuclear, Wilmington, DE. Sodium trichloroacetate was purchased from Aldrich Chemical Co., and recrystallized by the method of Murray (13). Nick translation kits were obtained from Amersham Corp., and Sequenase sequencing kits from United States Biochemical Co., Cleveland, OH. Nitrocellulose was purchased from Schleicher and Schuell, Keene, NH.

Genomic Library Screening—A library of oat genomic sequences in lambda phage Charon 34 was obtained from Dr. Mike Murray (Agri-genetics Corp., Madison, WI). For isolation of both globulin and avenin genomic clones, 0.9 – 1.2×10^6 phage were transferred to nitrocellulose filters as described by Maniatis *et al.* (14). Clones containing globulin gene sequences were identified by hybridization with a mixture of three DNA fragments labeled with ^{32} P by nick translation: a 350-bp *EcoRI*-*SalI* fragment isolated from the 5' end of oat globulin cDNA clone pOG2 (11); a 390-bp *XhoI*-*SpeI* fragment isolated from the 3' end of pOG2; and a 343-bp *Bam*HI-*DraI* fragment isolated from a globulin clone previously obtained from the same library by hybridization with oat globulin cDNA clone pOG77 (15). This latter genomic clone contains a missense mutation in its translation initiation codon.

Clones containing avenin gene sequences were identified by hybridization with a mixture of ^{32} P-labeled inserts from avenin cDNA

* The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

The nucleotide sequence(s) reported in this paper has been submitted to the GenBank™/EMBL Data Bank with accession number(s) J05485 and J05486.

‡ Recipient of National Science Foundation Fellowship in Plant Biology DCB-8608446.

§ Current address: Dept. of Biological Sciences, The University of Illinois at Chicago, P. O. Box 4348, Chicago, IL 60680.

¶ Recipient of Grant 87-2356 from the United States Department of Agriculture Competitive Research Grants Organization. To whom correspondence should be addressed: Dept. of Plant Sciences, 201 Forbes Hall, University of Arizona, Tucson, AZ 85721.

¹ The abbreviations used are: bp, base pair; kb, kilobase pair; SDS, sodium dodecyl sulfate.

clones pAV10, 21, 31, 41, and 122 (11).

Hybridizations were in $6 \times \text{SSC}$ ($1 \times \text{SSC} = 0.15 \text{ M NaCl}$, 0.015 M sodium citrate (pH 7.5)), 0.1% Ficoll, 0.1% polyvinylpyrrolidone, 0.1% bovine serum albumin, 0.01 M EDTA, 0.5% SDS, and 0.1 mg/ml sheared calf thymus DNA for 18 h at 65°C . The filters were washed four times for 15 min in $2 \times \text{SSC}$, 0.5% SDS at room temperature and four times for 15 min in $2 \times \text{SSC}$, 0.1% SDS at 65°C . Filters were then exposed to Kodak X-Omat film overnight at room temperature.

DNA Sequence Analysis—Globulin and avenin genomic clones were mapped with restriction endonucleases, and the restriction fragments containing gene sequences were identified by DNA gel blotting and hybridization with the cDNA fragments used in library screening by the procedures of Meinkoth and Wahl (16). These fragments were eluted from the gels, subcloned into the pGem-3Z plasmid (Promega, Madison, WI), and their sequences determined by the dideoxy chain termination method according to the instructions accompanying Sequenase sequencing kits. Fig. 1 shows the sequencing strategy used for one globulin and one avenin genomic subclone. For each subclone, 100% of the sequence was determined from both strands. DNA sequences were analyzed using the Microgenie computer program (Spinco Division, Beckman Instruments, Inc., Palo Alto, CA).

Transcript Mapping—RNA transcripts were mapped to the globulin and avenin genomic sequences by mung bean nuclease protection assays as described by Murray (13) except that the reaction mixture was extracted with phenol chloroform isoamyl alcohol (25:24:1) and precipitated with one-half volume of 7.5 M ammonium acetate and 2 volumes of ethanol. In these assays we used total RNA from membrane-bound polysomes isolated from developing oat seeds (8–10 days after anthesis) by the procedure of Larkins and Hurkman (17).

RESULTS

Isolation and Characterization of 12 S Globulin Genes—A library of oat genomic sequences in lambda phage Charon 34 was screened by hybridization with three DNA fragments, two derived from oat globulin cDNA clone pOG2 (11) and the third from a previously isolated globulin genomic clone that contains an inactive gene. Nine different clones were isolated and their restriction endonuclease sites mapped. DNA gel blot hybridizations indicated that each clone contained a single gene copy (data not shown). One clone, designated λOG1 , contained a complete gene sequence on a 3.3-kb *EcoRI* fragment. This fragment, OG1-E1, was subcloned into the plasmid pGem-3Z, and after additional rounds of subcloning of smaller fragments, its nucleotide sequence was determined (Fig. 2).

The fragment OG1-E1 contains a complete 12 S globulin coding sequence, as well as 982 bp of 5'-flanking sequence and 459 bp of 3'-flanking sequence. Comparison of this sequence with that of oat globulin cDNA clone pOG2 (11) reveals the presence of three introns of 117, 126, and 100 bp.

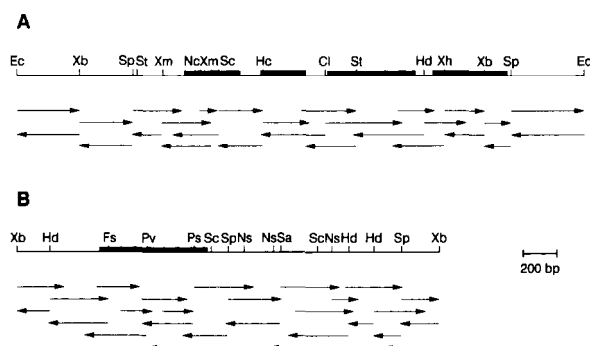


FIG. 1. Sequencing strategy for globulin genomic subclone OG1-E1 (A) and avenin genomic subclone AV45-X1 (B). Arrows indicate the direction and length of nucleotide sequence determinations. Restriction endonuclease sites are indicated as follows: Cl, *Clal*; Ec, *EcoRI*; Fs, *FspI*; Hc, *HincII*; Hd, *HindIII*; Nc, *NcoI*; Ns, *NsiI*; Pv, *PvuII*; Sc, *ScaI*; Sp, *SpeI*; St, *StuI*; Xb, *XbaI*; Xh, *XhoI*; Xm, *XmnI*.

These introns interrupt the coding sequence in the same positions in the coding sequence as the introns in genes for the storage globulins of soybean (18), pea (19), and other legumes (1), as well as the storage glutelin of rice (20, 21). Characteristic of introns in general, the three introns in OG1-E1 are AT-rich, consisting of 73% A + T compared to 54% A + T in the coding sequence. The splice sites of these introns conform to the GT/AG cleavage rule for eukaryotic genes (22).

Clone OG1-E1 encodes a polypeptide of 518 amino acids, including a 24-amino acid signal peptide, a 293-amino acid acidic polypeptide (M_r 33,112), and a 201-amino acid basic polypeptide (M_r 22,758). The mature polypeptide is rich in the amidated amino acids (21% Gln + Asn) and deficient in the sulfur-containing amino acids (2.0% Cys + Met), which is characteristic of seed storage globulins. The amino acid sequence of OG1-E1 has 97% identity to that encoded by oat globulin cDNA clone pOG2 (11). Like pOG2, OG1-E1 codes for four glutamine-rich octapeptide repeats near the COOH terminus of the acidic polypeptide (Fig. 2).

The 5' end of the mRNA was mapped by a mung bean nuclease protection assay (13) to a T residue 38 bp upstream of the ATG codon (Fig. 2). Potential TATAA and CCAT sequences are at positions -27 and -60 with respect to the start of the mRNA. There is one potential "CACA" element (23) at position -162 with the sequence 5'-CAACACAT-3'. The CACA element has been found in glycinin and β -conglycinin genes of soybean and has been shown to interact with embryo nuclear proteins.²

The 5'-flanking region of this gene lacks a number of conserved sequence elements that have been identified in other seed storage protein genes. These include the "legumin box" in pea legumin genes (24), the AACCCA element in soybean β -conglycinin genes (25, 26), the -300 element in several cereal prolamine genes (27, 28), and the CATGCATG RY element in many seed protein genes (29).

The translation initiation sequence in OG1-E1 is 5'-AAT-CATGGC-3'; the same sequence is found in oat globulin cDNA clones pOG2 and pOG3 (11) and genomic clone λOG7 (data not shown). This is a seven out of nine match with the consensus sequence for plant genes (5'-AACAAATGGC-3') reported by Lütcke *et al.* (30). The 3'-flanking region contains two 5'-AATAAA-3' polyadenylation sequences located 33 and 83 bp downstream of the TGA translation stop codon.

Isolation and Characterization of Avenin Genes—The oat genomic library was screened for avenin sequences by hybridization with inserts from four avenin cDNA clones (12). Out of about 1.2×10^6 phage clones, only four hybridizing phage were identified. Restriction endonuclease mapping showed that two of the four clones (λAV17 and λAV54) were identical and that the third (λAV45) was the same as those two except about 3.2 kb longer at its 5' end. The restriction map of the fourth clone (λAV53) was unrelated to the other three. Based on these results, clones λAV45 and λAV53 were characterized further.

DNA gel blotting showed that λAV45 contained all or parts of four avenin genes (Fig. 3) and that λAV53 contained only one coding sequence (data not shown). Preliminary sequence analysis revealed that the direction of transcription of all four genes on clone λAV45 is the same (Fig. 3). The presence of these four closely linked genes indicates that at least some avenin genes are tightly clustered in the oat genome.

Subcloning and DNA sequencing confirmed that clone λAV53 contains only one gene sequence. Detailed sequence analysis and comparisons with avenin cDNA sequences re-

² L. Pérez-Grau and R. B. Goldberg, personal communication.

-944 GAATCTTTT TGGAAAGTCA TTTTGCTCC TGAACCTCAG TGTTCCTAT TTAATAAAAA AACTAAAAA CTATATTAT AAGTTTGAAA AAATCATGAA
-844 ACAAATTTGT AAAAATTTGCT AGTGAATAT CCCACAAACG TGCAAAATCT CAATTGGAAG TGCTTTGTAT TTGAGCTAC ACAAATATGA CAAAGTGTGA
-744 CTTTTATGT GATTGAAAT CACTATACCT CAGATCTACA ATTTTGTCT TTTTGTGAAG CTATAAATAC ACATTATTTT GAGTTGAGAT TTTTCATGTT
-644 TGTGCTATGA ATCATAGGCT ACATCTGAT TTAATTTTAT AATTTTGTGG AACCTAAAAAT ATGTCTTAGA TTAATTTTAA AAAAGTGGGA TCATTGATGC
-544 CCATACACAC GAAATCTCCA CTCAATTTCT TTATACATTA TCTTTCTATA TCTACTAACG TGGATTATAC ATCATAGTAA GTTTCCTACT ACAATGCTTT
-444 TCTTTTACA ATGTGGACAT GACTCTTCCA CTTTTGGCT TTAATGTTGA TGGATATAC CATGACATGG AATTTTGTCC ACACACGTAG AACCATCCAT
-344 ATATATTGTT GTGCATAGAA CAAACACAA GCAAGCCATG AAAAGGAGT CACAGTGCC ACAAATGTT GTAGGAAGTA TAACATGAT GAGGCTTTT

-244 ATTGACGTC GGACAATGSC CAAGAGCTAC ATATAAAGA TGGTCTAGA TTTGTGATC AGCCACAGCT GTAGGCAGAA AACCAACAT ATCTTTGAG

-144 ACAAATTTAT ATCTATTAC ATTTAAACC ATGATCTGTT GAGTCACCAT GAATATCTTC TTAATCTATG TAATAATTAC ATGTGATCAT GTTTTCTCG

+1 MetAlaTh rThrArgPhe

-44 GACTACTTTT TATGGCTATA AAATCAAAT IACAATAAGG AAACCTAGCAC CAATCCACCT TCTACAATCT CTTCACAAAC TCATGGCAAC TACTAGGTTT

ProSerLeuL euPheTyrSe rTyrIlePhe LeuLeuCysA snGlySerMe tAlaGlnLeu PheGlyGlnS ePheThrPr oTrpGlnSer SerArgGlnG

+57 CCATCATGTT TGTITTAATC CTATATTTT CTCTGTGCA ATGGGTCAAT GGCTCAGCTA TTGGGGCAGA GCITTAACCC ATGGCAAGC TCTGACAAAG

lyGlyLeuAr gGlyCysArg PheAspArgL euGlnAlaPh eGluProLeu ArgGlnValG lySerGlnAl eGlyIleIle GluTyrPheA spGluGlnAs

+157 GAGGTTTAA GGGGTGAGA TTTGATAGC TACAAGCATT TGAACCACTT CGACAAGTGG GGTCAACAGC GGTATCAATT GAGTACTTTG ATGAGCAGAA

lGluGlnPhe ArgCysAlaG lyValSerVa lIleArgArg ValIleGluP roGlnGlyLe uLeuProGln GlnTyrHisA snAlaProGl yLeuValTyr

+257 TGAGCAATTT CGTGTGAGC GTGTATCGGT CATCGTGTGT GTTATTGAGC CTCAGGCGCT CTGTGTACCT CAATACCACA AGCTCTCTGG CTGTGTGTAC

IleLeuGlnG

+357 ATCCTTCAAG GTTAGTGTCT AATTGATTAT AAAAATGGC TTTGTATAC TTCACITAGG ATTATGATGT GCCAAATGTT ACACCGTTCA TATTTTTTAA

lyA rgGlyPheTh rGlyLeuThr PheProGlyG ysProAlaTh rPheGlnGln GlnPheGlnP roPheAspGl

+457 CAATGAACA AATGTGTTTT CTTTATAGTA GGGGATTCAC AGGGTTGACT TTCCCTGGAT GCCCGGCGAC CTTCACAAAC CAGTTCCAAC CAITTTGATCA

nAlaGlnPhe AlaGluGlyG lnSerGlnSe rGlnAsnLeu LysAspGluB isGlnArgVa lHisIleIle LysGlnGlyA spValValAl aLeuProAla

+557 AGCCCAAGTT GCTGAAGGTC AAAGCAAAG CCAAAATCTT AAGGATGAAC ACCAAGAGT TCACCACATC AAACAAGAGT ATGTTGTGTC TCTACCGGCT

GlyIleValH isTrpCysTy rAsnAspGly AspAlaProI leValAlaVa lTyrValPhe AspValAsnA snAsnAlaAs nGlnLeuGlu ProArgGlnL

+657 GGCAATAGTA ACTGTGTCTA CAACGATGGT GATGACCGGA TTGTAGCTGT CTATGTCTTC GACGTAAACA ACAACGCTAA TCAGCTTGAA CCAAGGCAAA

ys

+757 AGGTAACAT ACACTTAAT GCACACAAA TATATAACAC TATTTACAAG TGATCCAACAT ATGATTGTGA GTATACATTA GTGGGATATT AATGAACCT

Gl uPheLeuLeu AlaGlyAsnA snLysArgGl uGlnGlnPhe GlyGlnAsnI lePheSerGl yPheSerVal

+857 GTTAACTTC ATCGATATAA AATTITAGGA GTCTTGTG GCTGGTAAAC ACAAGAGAGA GCAACAGTTT GGACAAAAA TATTCAGTGG ATTCAAGTTC

GlnLeuLeuS eGluAlaLe uGlyIleSer GlnGlnValA leGlnLysI l eGlnSerGln AsnAspGlnA rgGlyGluI l eIleArgVal SerGlnGlyL

+957 CAACTTCTTA GTGAGGCCCT TGGTATAAGT CAGCAAGTAG CACAAAAGAT TCAGAGTCAA AATGACCAAA GAGTGAGAT AATTCGTGTC AGTCAAGGCC

euGlnPheLe uLysProPhe ValSerGlnG lnGlyProVa lGluHisGln AlaTyrGlnP roIleGlnSe rGlnGluGlu GlnSerThrG lnTyrGlnVa

+1057 TTCAATTTCTT GAAGCCTTTT GTTTCACAC AAGGACCACT AGAGCATCAA GCCTCAACAC CAATTCAAAG TCAAGAGAA CAATCAACCT AATACCAAGT

lGlyGlnSer ProGlnTyrG lnGluGlyGl nSerThrGln TyrGlnProG lyGlnSerTr AspGlnSer PheAsnGlyL euGluGluA nPheCysSer

+1157 AGGCAATCA CCAATATC AAGAAGGACA ATCACTCAA TACCAGCCAG GACAGTCATG CBACCAAGT TTCAATGTTT TGGAGAGAA TTTCTGTCTA

LeuGluAlaA rgGlnAsnI l eGluAsnPro LysArgAlaA spThrTyrAs nProArgAla GlyArgIleT hrHisLeuAs nSerLysAsn PheProThrL

+1257 TTGAGGCAA GGCAAAACAT CGAAAACCCG AAAGGTGCGC ACACGTACAA CCCACGTGCT GCGAGATAA CACATCTCAA TACGAAGAT TTTCCACCC

euAsnLeuVa lGlnMetSer AlaThrArgV aIAsnLeuTy rGln

+1357 TTAACCTGCT GCAATGAGT GCTACAAGAG TAAATTTATA CCAGGTATTT ATGATACTAC ATTCAACACA CTATCTTATT TTTAGATATT CTAAGCTTCA

AsnAla IleLeuSerP roTyrTrpAs nIleAsnAla HisSerValM etHisMetI l

+1457 TACAACCGAT TAAATATATG GCATACAAAT AATTGCTATT GCAGAATGCT ATTCCTTCAC CATCTCGGAA CATTATGCT CACAGTGCA TGCACATGAT

eGlnGlyArg AlaArgValG lnValValAs nHisHisGly GlnThrValP heAsnAspI l eLeuArgArg GlyGlnLeuL euIleIlePr oGlnHisTyr

+1557 CCAAGGACGT GCTCGAGTTC AAGTTGTCAA TAACCATGCT CAGACCGTAT TCAATGACAT TCTTCGTCCG GGACAACTGC TAATCATACC ACAACACTAT

ValValLeuL ysAlaGluI uArgGluGly CysGlnTyrI leSerPheLy sThrAsnPro AsnSerMetV alSerGlnI l eAlaGlyLys ThrSerIleL

+1657 GTTGTCTCA AGAAGGCAGA GCGTGAAGGA TGCCAGTATA TTTCAATCAA GACCAACCCC AATTCCATGG TTAGCCAGAT CGCAGGAAG ACTTCCATCC

euArgAlaLe uProValAsp ValLeuAlaA snAlaTyrAr gIleSerArg GlnGluAlaG lnAsnLeuLy sAsnAsnArg GlyGluGluP heAspAlaPh

+1757 TACGTGCAAT GCCCGTTGAT GTCTCGCCA ATGATACCGC CATTTCTAGA CAGGAAGGCC AAAACCTCAA AAATAACCGT GGAGAAGAGT TTGATGCATT

eThrProLys PheThrGlnT hrGlySerGl nSerTyrGln AspGluGlyG luSerSerSe rThrGluLys AlaSerGluE nd

+1857 CACCCCTAAG TTTACACAAA CGGCTCCCA GAGTTACAG GACGAGGGGG AGTCATCTTC GACTGAGAAG GCATCCGAGT GAATAGTGA GTGTAAATGA

+1957 AACTAGTATA GTGAAATAA GGCAATCGAT GTTTCAGCC TAGTGTATA TAACCGCTTA TCTCAATAA AAAGTTTCTC CGTGTATAT TTTTGTCTTG

+2057 TTTCTGTGAT TCTTCTTAAT TTTATCTTTT ATGATCAACC AACTTACCTC TCTTCTTCAA TTTTCTGCA CCITTCACAT ATGCATATCA AGGAGACGTA

+2157 TAGTCCACTT TACTAGATT TATTTTGTGG TACTATCTTT TGTGGTTTGA GTATTTCTTA GTTCCGATAA AAACATCTAA TTTACTGATG CTAATTTTTT

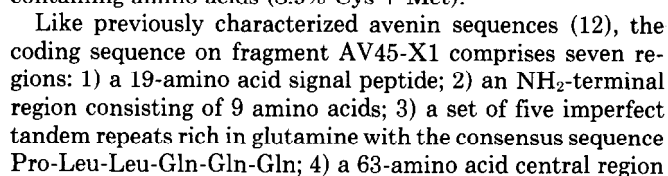
+2257 TCAGCAAAAC TAGCCACCTA AAATTTGTGA TCACATACCT TCAATTTCT ATGCCAAAGT GTGCGATTTT TATGTCTAT AGATCCAAGC TTGGGCTGCA

+2357 GGTGAGCTCT AGAGGATCGA TCCCGGGGCG AGCTCGAATT C

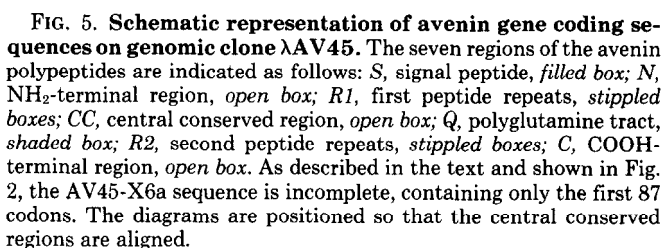
FIG. 2. Nucleotide and deduced amino acid sequence of oat globulin genomic subclone OG1-E1. Shown is the complete sequence of a 3341-bp *Eco*RI fragment isolated from genomic clone λ OG1. Potential TATAA and CCAT sequences are underlined as are AATAAA polyadenylation sequences. The 5' end of the mRNA was mapped to the T residue numbered +1. A potential CACA sequence at position -162 is overlined. The presumed site of signal peptide cleavage is indicated by a solid arrowhead. The site of proteolytic processing into acidic and basic polypeptides is indicated by an open arrowhead. The cysteine residues involved in interchain disulfide bond formation are double-underlined. The octapeptide repeats at the COOH terminus of the acidic polypeptide are boxed.

vealed that the coding sequence of λ AV53 contains four in-frame translation stop codons, seven single-base insertions or deletions, and a central deletion of about 100 bp. Moreover, the 5' noncoding region does not possess TATAA or CCAT

sequences or any of the conserved sequence elements found in cereal prolamine genes (see later discussion). We concluded that the gene on clone λ AV53 is inactive and did not characterize it further.



Avenin genes contain inverted repeat sequences located in different regions of the coding sequence. In AV45-X1 there is a series of overlapping inverted repeats in the coding sequence



for the first set of peptide repeats. Four of these inverted repeats are shown in hairpin loop configurations in Fig. 6. The computer-calculated free energy change of the most energetically favored of these hairpins is -26.0 kcal/mol. The base pairing in these hairpins is largely the consequence of the spacing of TTG leucine codons and CAA glutamine codons. Inverted repeats are also present in the first peptide repeat regions of AV45-X3a and AV45-X6a, the other members of the avenin gene subfamily that contains AV45-X1 (data not shown). Members of the other avenin gene subfamily also contain inverted sequence repeats. Both genomic subclone AV45-X2 and cDNA clone pAV10 (12) have a 9-bp inverted repeat in the signal peptide coding region. The computer-calculated free energy change for hairpin formation in AV45-X2 is -17.2 kcal/mol, and in pAV10 it is -24.4 kcal/mol.

Comparison between Globulin and Avenin Genes—Although oat globulin and avenin genes are both highly expressed in endosperm tissue at the same stages of development, sequence comparisons of their promoter regions failed to identify conserved sequence elements that might determine tissue specificity and developmental timing. Furthermore, the coding regions of the two classes of genes have distinctly different codon usage patterns (Table I). Globulin genes show a slight but significant bias towards A (28%) and T (29%) residues in the third position of the codons, but avenin genes exhibit a marked preference for C (30%) and especially G (43%) residues. For example, 25 of the 33 codons for alanine in OG1-E1 are GCT or GCA, but 13 of the 14 alanine codons in AV45-X1 are GCC or GCG. For glutamine, the most numerous

TABLE I
Codon distribution of oat globulin gene OG1-E1 and avenin gene AV45-X1

OG1-E1			AV45-X1		OG1-E1			AV45-X1	
Phe	TTT	15	2	Tyr	TAT	5	1		
	TTC	15	10		TAC	13	2		
Leu	TTA	3	0	End	TAA	0	1		
	TTG	12	8		TAG	0	0		
	CTT	11	2	His	CAT	3	1		
	CTC	6	6		CAC	8	1		
	CTA	5	1	Gln	CAA	50	25		
	CTG	2	5		CAG	19	34		
Ile	ATT	12	2	Asn	AAT	17	0		
	ATC	9	8		AAC	17	2		
	ATA	6	0	Lys	AAA	3	0		
	ATG	6	12		AAG	12	1		
Val	GTT	11	1	Asp	GAT	9	1		
	GTC	7	4		GAC	6	1		
	GTA	9	1	Glu	GAA	11	1		
	GTG	4	6		GAG	18	4		
Ser	TCT	3	0	Cys	TGT	2	0		
	TCC	7	1		TGC	5	8		
	TCA	11	1	End	TGA	1	0		
	TCG	1	1	Trp	TGG	4	0		
Pro	CCT	6	5	Arg	CGT	11	0		
	CCC	3	2		CGC	2	0		
	CCA	12	5		CGA	3	1		
	CCG	4	5		CGG	0	3		
Thr	ACT	5	2	Ser	AGT	11	1		
	ACC	8	3		AGC	6	5		
	ACA	4	2	Arg	AGA	6	0		
	ACG	2	0		AGG	7	3		
Ala	GCT	13	0	Gly	GGT	11	1		
	GCC	6	6		GGC	6	3		
	GCA	12	1		GGA	14	0		
	GCG	2	7		GGG	7	1		

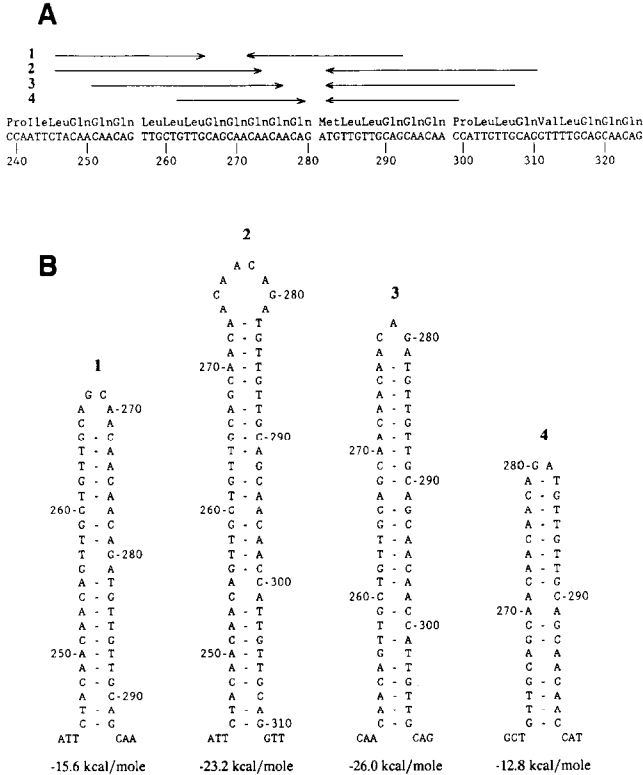


FIG. 6. Potential hairpin loop structures in avenin gene AV45-X1. A, the nucleotide and amino acid sequence of the first repeat region of gene AV45-X1. Shown are nucleotides 239 through 325 from Fig. 4. Overlapping inverted sequence repeats are numbered 1 through 4 and are indicated by arrows. B, inverted repeats 1 through 4 arranged in hairpin loop configurations. The nucleotides are numbered as in A. Computer-calculated free energy changes for hairpin formation are given below each sequence.

amino acid in each sequence, 50 of the 69 codons in OG1-E1 are CAA. In AV45-X1, however, the 59 codons are slightly biased toward CAG.

DISCUSSION

We have isolated and characterized genomic clones corresponding to the two major classes of seed storage proteins in oats, the 12 S globulins and the avenins, as a step toward understanding the regulation of the synthesis of these proteins during oat seed development. Our previous experiments (12) showed that although mature oat endosperm contains 75–80% globulin by weight and only 10–20% avenin (2, 3), mRNA levels for the two protein classes are roughly equal throughout development. In the current study we have analyzed globulin and avenin gene sequences in order to gain some insight into mechanisms of regulation of these storage protein genes.

We isolated several 12 S globulin genes and found that their coding sequences are interrupted by three short introns, two in the acidic polypeptide coding region and one in the basic polypeptide coding region (Fig. 2). The positions of the introns in the oat globulin genes are the same as in genes for other seed storage globulins, both in dicots (1) and in another monocot, rice (20, 21). The conservation of gene structure in these evolutionarily diverse plants is consistent with the idea that the storage globulin genes existed before the division of angiosperms into monocots and dicots (31). It further implies that the globulins are the more primitive seed storage proteins and that the prolamines arose in the monocots after their separation from the dicots. The ancestral storage globulin gene presumably contained three introns, the first of which was lost from legumin genes of *Vicia faba* (24) and the third from 11 S globulin genes of sunflower (32).

We mapped the 5' end of the globulin mRNA to a T residue

38 bp upstream of the ATG codon in clone OG1-E1 (Fig. 2). This would give a 5' noncoding sequence that corresponds in length with several previously characterized oat 12 S globulin cDNA clones (11). The oat globulin cDNA sequence that we published contains 145 bp of 5' noncoding sequence, however (11). We have since found that the published sequence contains a 5'-terminal inverted repeat, with nucleotides 10–110 an inverted repeat of nucleotides 210–310. This is a probable artifact of cloning.

The 5' noncoding region of OG1-E1 lacks several conserved sequence elements found in other storage globulin genes that have been suggested to be involved in the tissue specificity or developmental regulation of gene expression. These putative regulatory sequences include the legumin box of pea legumin and other legume 11 S globulins (24), the AACCCA element of soybean β -conglycinins (25, 26), and the CATGCATG RY element of many dicot seed proteins (29). A second oat globulin genomic clone, λ OG7, also lacks these sequences (data not shown).

It is not surprising that oat globulin genes do not possess the same conserved upstream regulatory sequences as analogous genes in dicots, since the oat genes are expressed in endosperm tissue and the dicot genes in the storage parenchyma of the cotyledons. If monocots evolved from an ancient dicot, as is commonly assumed, the regulation of the storage globulin genes in the monocot progenitor of oats and rice changed to give high expression in endosperm tissue. Similar regulatory sequences may have eventually been incorporated into the prolamine genes. Alternatively, globulin genes may have come under the control of an endosperm-specific prolamine promoter. This could have involved the evolution of regulatory sequences that direct high levels of endosperm-specific gene expression. Detailed promoter expression studies will be required to positively identify the regulatory elements in the oat globulin genes.

Avenin clones proved to be difficult to isolate from the oat genomic library, unlike the readily obtained globulin clones. In the present study, the screening of 1.2×10^6 phage clones yielded only two non-colinear avenin genomic clones, λ AV53 and λ AV45. Clone λ AV53 contained a single gene copy whose sequence included numerous in-frame nonsense and missense mutations as well as one large deletion. This gene also lacked the expected conserved sequences in its promoter region and is unquestionably transcriptionally inactive. The second clone, λ AV45, contained four genes on four different *Xba*I fragments: 1) AV45-X2 contained a complete gene with an intact promoter region but with a stop codon at position 44 of the coding sequence; 2 and 3) AV45-X3a and AV45-X1 each contained a complete promoter region and coding sequence; and 4) 45-X6a was at the 5' end of the insert and contained a partial gene sequence possessing only the first 87 codons of the coding sequence (Fig. 3).

Prolamine genes with coding sequences interrupted by in-frame stop codons have been isolated from maize (33), wheat (34, 35), and barley (36). In addition, we previously obtained an oat globulin genomic clone with a missense mutation at the translation initiation codon. Thus it is not surprising to find inactive genes in the avenin multigene family, which consists of about 25 genes (11, 12). Nevertheless, it is unlikely that the high proportion of inactive avenin genes we recovered from the genomic library (two out of five) is an accurate reflection of the number of inactive genes in the gene family. This bias towards isolation of clones containing inactive genes from the genomic library may instead be a consequence of the selective loss of phage clones with active genes during library amplification.

Previously, we recovered disproportionately few avenin cDNA clones from a λ gt11 expression library relative to the high levels of avenin mRNA that are present in oat endosperm (12). We suspect that in this case the low recovery could have resulted from toxicity of avenin polypeptides to the bacteria used for plating. Such toxicity has been observed with the α -zeins of maize.³

We found the structure of avenin genes on clone λ AV45 to be typical of monocot prolamine genes in that they contain no introns (Fig. 4) (11). The promoter regions of the avenin genes contained expected TATAA and CCAT sequences as well as –300 elements at positions –311 or –333 with respect to the ATG codon. The –300 element has been hypothesized to be involved in the tissue-specific or developmental regulation of cereal prolamine genes (28). In three of the four oat avenin genes, the –300 element has the sequence 5'-TTTGA-CATGTAAAGTGAAT-3', a 17 of 19 match with the consensus sequence in prolamine genes of maize, wheat, and barley (28). The sequence in AV45-X1 is 5'-ATTGACATGTAA-AGCGAAA-3', a 15 of 19 match. The –300 conserved sequences are present in the promoters of oat avenin genes but not globulin genes, despite the much higher synthesis of globulins than avenins in oat endosperm. Globulin genes must contain other regulatory sequences that direct high rates of transcription in endosperm. These findings also suggest that if the –300 element is responsible for tissue-specific transcription of avenin genes in oats, it is insufficient to insure high levels of protein synthesis.

The presence of four gene sequences on clone λ AV45 indicates that at least some avenin genes are tightly clustered in the oat genome. Moreover, of these four genes, one belongs to the α gene subfamily and three to the γ gene subfamily, demonstrating that genes from the two subfamilies may be closely linked. This conclusion is consistent with previous genomic DNA gel blots (12), in which some hybridizing bands contained multiple gene copies. Close linkage of prolamine genes has also been observed in maize (37, 38) and rice (39).

We previously showed that there are roughly equal amounts of globulin and avenin mRNAs throughout the development of oat endosperm (12), despite there being five to seven times as much globulin protein as avenin in mature seeds (2, 3). Others have shown that only 25–30% of the total polysomal mRNA in developing oat endosperm is globulin mRNA (40), but that translation of this polysomal mRNA *in vitro* yields predominantly globulin polypeptides (41). These results suggest that the control of storage protein synthesis in oats is complex and involves regulation not only at transcription but also at translation.

The gene sequences presented in the current study provide some clues as to why equal amounts of globulin and avenin mRNAs do not give rise to equal amounts of proteins in oat endosperm. First, the sequences at the translation initiation site of the two classes of genes differ. The globulin sequence (5'-AATCATGGC-3'; Fig. 2) is a closer match with the consensus for plant genes (5'-ACAATGGC-3') (30) than are the avenin sequences (5'-CACCATGAA-3' and 5'-TACCATGAA-3'; Fig. 4). Adenines are found at positions +4 and +5 of the avenin sequence, which is rare in plant genes but rather common in animal genes. Interestingly, the avenin sequence more closely resembles the animal consensus (5'-ACCATGC-3') (42). Lütcke *et al.* (30) suggested that the strong preference for G at position +4 and C at position +5 in plant genes indicates that nucleotides +4 and +5 may modulate initiation codon selection in plants as has been shown for nucleotide

³ R. E. Cuellar, D. W. Galbraith, and B. A. Larkins, unpublished observation.

-3 in animals (43). If so, the initiation of translation of globulin mRNAs may be more efficient than avenin mRNAs, which have A residues at positions +4 and +5. It remains to be determined whether these sequence differences in fact influence rates of translation initiation in oat endosperm.

Second, the potential hairpin loop structures that we found in the coding sequences in both subfamilies of avenin genes (Fig. 6) might reduce the rate of translation of avenin mRNAs. Hairpin structures in zein mRNAs have been shown to affect their translation (44), but it has not been demonstrated that these hairpins occur naturally in zein genes or whether they are merely cloning artifacts (45). In the avenin genes, it is uncertain whether the stability of the hairpins would be high enough to interfere with translation.

Third, the difference in codon usage between globulin and avenin genes, with a preference for C and G residues in the third codon positions in the avenins (Table I), may also affect the rate of translation of the respective mRNAs. Similar differences in codon usage patterns have been observed in the α -, β -, and γ -zein genes in maize (46-48). It is possible that the tRNA population in developing oat endosperm is more closely correlated with the codon distribution of the globulin genes than of the avenin genes. As a result, avenin mRNAs may be translated more slowly than globulin mRNAs. One mechanism by which avenin mRNAs might be more slowly translated is suggested by the coding sequences of the signal peptides of the avenins. Fig. 4 shows that the signal peptide of avenin gene AV45-X1 contains 5 closely spaced alanine residues, all encoded by either GCC or GCG. The three other avenin gene sequences and the three avenin cDNA sequences reported earlier (12) contain 5-7 alanines in their signal peptides, and all of them are also encoded by GCC or GCG. These two codons are relatively infrequent in globulin coding sequences (Table I). Thus, if tRNA_{GCC} or tRNA_{GCG} or both are in low abundance in oat endosperm, it could cause stalling of ribosomes on the avenin mRNAs shortly after translation initiation and result in reduced rates of translation. Whether different tRNA populations in fact exist in oat endosperm has yet to be established, however.

Finally, our isolation of such a high percentage of inactive avenin gene sequences leads to the question of whether some of the mRNAs we have measured in developing endosperm are likewise incapable of giving rise to avenin polypeptides. It is unlikely, however, that sufficiently large amounts of inactive avenin mRNAs could be present in developing oat endosperm to explain our results.

In summary, it appears that the avenin genes of oats have evolved regulatory sequences that give high rates of transcription, but that globulin mRNAs are preferentially translated in oat endosperm. This disparity in translational efficiencies results in a high-globulin, low-prolamine protein composition in the oat seed. Experiments are underway to elucidate the unusual regulation of globulin and avenin gene expression in oats.

Acknowledgment—We acknowledge the contribution of A. Lee Lang in the isolation of globulin genomic clones.

REFERENCES

- Shotwell, M. A., and Larkins, B. A. (1989) in *The Biochemistry of Plants, A Comprehensive Treatise* (Marcus, A., ed) Vol. 15, pp. 297-345, Academic Press, San Diego, CA.
- Peterson, D. M., and Smith, D. (1976) *Crop Sci.* **16**, 67-71.
- Colyer, T. E., and Luthe, D. S. (1984) *Plant Physiol.* **74**, 455-456.
- Brinegar, A. C., and Peterson, D. M. (1982) *Arch. Biochem. Biophys.* **219**, 71-79.
- Peterson, D. M. (1978) *Plant Physiol.* **62**, 506-509.
- Walburg, G., and Larkins, B. A. (1983) *Plant Physiol.* **72**, 161-165.
- Kim, S. I., Charbonnier, L., and Mossé, J. (1978) *Biochim. Biophys. Acta* **537**, 22-30.
- Bietz, J. A. (1982) *Biochem. Genet.* **20**, 1039-1053.
- Pernollet, J.-C., Huet, J.-C., Galle, A.-M., and Sallantin, M. (1987) *Biochimie* **69**, 683-689.
- Egorov, T. A. (1988) *Bioorg. Khim.* **14**, 959-962.
- Shotwell, M. A., Afonso, C., Davies, E., Chesnut, R. S., and Larkins, B. A. (1988) *Plant Physiol.* **87**, 698-704.
- Chesnut, R. S., Shotwell, M. A., Boyer, S. K., and Larkins, B. A. (1989) *Plant Cell* **1**, 913-924.
- Murray, M. G. (1986) *Anal. Biochem.* **158**, 165-170.
- Maniatis, T., Fritsch, E. F., and Sambrook, J. (1982) *Molecular Cloning: A Laboratory Manual*, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.
- Walburg, G., and Larkins, B. A. (1986) *Plant Mol. Biol.* **6**, 161-169.
- Meinkoth, J., and Wahl, G. (1984) *Anal. Biochem.* **139**, 267-284.
- Larkins, B. A., and Hurkman, W. J. (1978) *Plant Physiol.* **62**, 256-263.
- Nielsen, N. C., Dickinson, C. D., Cho, T.-J., Thanh, V. H., Scallan, B. J., Fischer, R. L., Sims, T. L., Drews, G. N., and Goldberg, R. B. (1989) *Plant Cell* **1**, 313-328.
- Lycett, G. W., Croy, R. R. D., Shirsat, A. H., and Boulter, D. (1984) *Nucleic Acids Res.* **12**, 4493-4506.
- Takaiwa, F., Ebinuma, H., Kikuchi, S., and Oono, K. (1987) *FEBS Lett.* **221**, 43-47.
- Okita, T. W., Hwang, Y. S., Hnilo, J., Kim, W. T., Aryan, A. P., Larsen, R., and Krishnan, H. B. (1989) *J. Biol. Chem.* **264**, 12573-12581.
- Breathnach, R., Benoist, C., O'Hare, K., Gannon, F., and Chambon, P. (1978) *Proc. Natl. Acad. Sci. U. S. A.* **75**, 4853-4857.
- Goldberg, R. B. (1986) *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **314**, 343-353.
- Bäumlein, H., Wobus, U., Pustell, J., and Kafatos, F. (1986) *Nucleic Acids Res.* **14**, 2707-2720.
- Chen, Z.-L., Schuler, M. A., and Beachy, R. N. (1986) *Proc. Natl. Acad. Sci. U. S. A.* **83**, 8560-8564.
- Allen, R. D., Bernier, F., Lessard, P. A., and Beachy, R. N. (1989) *Plant Cell* **1**, 623-631.
- Kreis, M., Shewry, P. R., Forde, B. G., Forde, J., and Milfin, B. J. (1985) in *Oxford Surveys of Plant Molecular and Cell Physiology* (Milfin, B. J., ed) Vol. 2, pp. 253-317, Oxford University Press, Oxford.
- Thompson, G. A., and Larkins, B. A. (1989) *BioEssays* **10**, 108-113.
- Dickinson, C. D., Evans, R. P., and Nielsen, N. C. (1988) *Nucleic Acids Res.* **16**, 371.
- Lütcke, H. A., Chow, K. C., Mickel, F. S., Moss, K. A., Kern, H. F., and Scheele, G. A. (1987) *EMBO J.* **6**, 43-48.
- Boroto, K., and Dure, L., III (1987) *Plant Mol. Biol.* **8**, 113-131.
- Vonder Haar, R. A., Allen, R. A., Cohen, E. A., Nessler, C. L., and Thomas, T. L. (1988) *Gene (Amst.)* **74**, 433-443.
- Kridl, J. C., Vieira, J., Rubenstein, I., and Messing, J. (1984) *Gene (Amst.)* **28**, 113-118.
- Forde, J., Malpica, J.-M., Halford, N. G., Shewry, P. R., Anderson, O. D., Greene, F. C., and Milfin, B. J. (1985) *Nucleic Acids Res.* **13**, 6817-6832.
- Rafalski, J. A. (1986) *Gene (Amst.)* **43**, 221-229.
- Entwistle, J. (1988) *Carlsberg Res. Commun.* **53**, 247-258.
- Spena, A., Viotti, A., and Pirrotta, V. (1983) *J. Mol. Biol.* **169**, 799-811.
- Kriz, A. L., Boston, R. S., and Larkins, B. A. (1987) *Mol. Gen. Genet.* **207**, 90-98.
- Kim, W. T., and Okita, T. W. (1988) *Plant Physiol.* **88**, 649-655.
- Fabijanski, S., Matlashewski, G. J., and Altosaar, I. (1985) *Plant Mol. Biol.* **4**, 205-210.
- Fabijanski, S., and Altosaar, I. (1985) *Plant Mol. Biol.* **4**, 211-218.
- Kozak, M. (1984) *Nucleic Acids Res.* **12**, 857-872.
- Kozak, M. (1986) *Cell* **44**, 283-292.
- Spena, A., Krause, E., and Dobberstein, B. (1985) *EMBO J.* **4**, 2153-2158.
- Galili, G., Kawata, E. E., Cuellar, R. E., Smith, L. D., and Larkins, B. A. (1986) *Nucleic Acids Res.* **14**, 1511-1524.
- Geraghty, D. E., Messing, J., and Rubenstein, I. (1982) *EMBO J.* **1**, 1329-1335.
- Pedersen, K., Argos, P., Naravana, S. V. L., and Larkins, B. A. (1986) *J. Biol. Chem.* **261**, 6279-6284.
- Prat, S., Cortadas, J., Puigdomènech, P., and Palau, J. (1985) *Nucleic Acids Res.* **13**, 1493-1504.